The Dissertation Committee for Giorgio Paulon
certifies that this is the approved version of the following dissertation:

Bayesian Partition Models for Local Inference in Longitudinal and Survival Data

Committee:

Abhra Sarkar, Supervisor

Peter Müller, Co-Supervisor

Mingyuan Zhou

Bharath Chandrasekaran

Bayesian Partition Models for Local Inference in Longitudinal and Survival Data

by

Giorgio Paulon

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2021

A Lucia, Gianni e Anna, senza i quali non sarei arrivato qui.

Acknowledgments

Many people, knowingly or not, helped me during my doctoral studies. While it would be impossible to mention every single one of them, I am grateful to each person who shaped me and my research experience during these past five years.

First and foremost, I am extremely thankful to my supervisors Dr. Abhra Sarkar and Prof. Peter Müller for their mentoring and guidance. Your thought-provoking advice, unshaken trust and constant encouragements were determining factors in my personal and professional development. I could not have asked for better role models and I feel extremely fortunate to have you as my supervisors.

Thanks to all the people who worked with me and directly influenced my research experience during these five years: Prof. Bharath Chandrasekaran, Dr. Rachel Reetzke, Dr. Fernando Llanos, Dr. Casey Roark and Prof. Victor Sal y Rosas. I would also like to thank Dr. Mingyuan Zhou for serving as my committee member as well as the Department of Statistics and Data Sciences for the financial support and the opportunity of teaching as an assistant instructor.

I would not be able to celebrate this milestone if it was not for my family and friends in Italy: being far apart during these years has been challenging, but I know that I can find home anytime I am with you. To my friends and colleagues in Austin, your companionship and support were crucial for me. To Alfredo, thank you for always being at my side and for being such a loving and compassionate person.

Bayesian Partition Models for Local Inference in Longitudinal and Survival Data

Publication No. _____

Giorgio Paulon, Ph.D.
The University of Texas at Austin, 2021

Supervisors: Abhra Sarkar
Peter Müller

This dissertation proposes novel Bayesian semiparametric and nonparametric methods for complex, large and potentially high-dimensional longitudinal and survival data. The first part, comprising the bulk of this thesis, develops sophisticated dynamic partition models for longitudinal data that allow common features to be shared across some time segments while differing across others. These ideas are then specifically adapted to develop novel drift-diffusion models for the analysis of behavioral data on category learning in auditory neuroscience. The second part of this work proposes a bivariate survival regression method, borrowing information across two outcomes via common features in parts of the induced marginal partitions. In terms of flexibility and interpretability, the methods presented here provide significant improvements over many previously available tools and techniques, leading to interesting, novel and meaningful inference in many diverse application areas.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

This chapter illustrates the research motivations underlying my doctoral thesis. In the following, we focus on the development of novel statistical approaches that aid in the study of complex phenomena and provide new insights into related scientific queries. The motivating applications arise from diverse areas, including auditory neuroscience and disease recurrence. We believe that the contributions presented in this thesis have the potential to improve current scientific results and statistical practice. A brief overview of our methodological innovations are presented, followed by the general thesis outline.

1.1 Scientific Motivations

The line of research presented in Chapter 3 is motivated by auditory behavioral neuroscience experiments studying Mandarin tone learning mechanisms in non-native speakers. The longitudinal evolution of the underlying mechanisms during the learning period is critical to understand the cognitive dynamics of speech learning. Neuroscientists are usually interested in assessing if and how the perceptual stimuli affect the underlying learning processes at different longitudinal stages of the experiments. The statistical challenge is to make this assessment from data

on response accuracies and response times.

Chapter 4 proposes a novel framework for regression of longitudinal data with high dimensional categorical predictors. When dealing with highly heterogeneous data collected over time, assessing the local variations in the response profiles can provide valuable insights into the underlying data generating mechanisms. However, most existing methods only allow for curve profiles that are either completely different or completely identical across the entire longitudinal domain. Such models can be hard to interpret and they have small estimation efficiency when applied to data that exhibit 'local' heterogeneity.

Chapter 5 considers the open problem of estimating a joint distribution of two outcomes - specifically, infection time and time to symptoms - from current status data of infectious diseases. This particular research is motivated by a randomized controlled trial studying recurrent infections of sexually transmitted diseases such as chlamydia and gonorrhea. Current status datasets are routinely encountered in many medical applications, i.e. when patients fill a symptoms survey and are tested for a specific disease. Despite extensive literature on inference approaches for current status data, there are no widely available methods for the case of bivariate outcomes.

1.2 Research Contributions

In Chapter 3, we identify the tone learning problem as a decision-making task under perceptual stimuli. Drift-diffusion processes are popular models for decision-making as they can mimic evidence accumulation through increased firings of neurons in the brain. The statistical challenge then translates into developing a flexible

locally varying longitudinal drift-diffusion mixed model. The proposed model and associated computational machinery make innovative use of B-spline mixtures, hidden Markov models (HMM) etc. to address daunting statistical challenges. Applied to the motivating longitudinal tone learning study, the proposed method provides significant novel insights into the cognitive dynamics, allowing us to answer important scientific questions outside the scope of the previously existing literature. These include a detailed understanding of how biologically interpretable model parameters evolve with learning, differ between input-response tone combinations, and differ between well and poorly performing adults. The novel drift-diffusion model has also been used in further studies (Roark *et al.*, 2021) to understand how learning across modalities (i.e. auditory and visual) can share some characteristics.

Chapter 4 generalizes some of the previous ideas to the setting of multiple predictors. In particular, we propose an efficient and flexible Bayesian semiparametric longitudinal mixed model for functional data in the presence of multiple categorical covariates. The proposed model infers dependent random partitions of the covariate space at several locations. These local partitions allow for the automated assessment of the predictors' time-varying local influences on the response variable. Additionally, shrinkage priors on the partition structures enable dynamic variable selection by eliminating redundant predictors, which in turn facilitates model interpretability. Although the literature on longitudinal data analysis methods is vast, such problems had not been addressed before.

In Chapter 5, we address the nonidentifiability issues of nonparametric estimation of a bivariate density for current status data. We show that simple assumptions

on the dependence structure between the two events can fix this problem, and we design a model that is tailored to the motivating case study. Additionally, we move beyond the independent censoring assumption by exploiting some minimal known dependence between censoring times and event times. Applied to a recurrent infection study, the method provides novel insights into how symptoms-related hospital visits are affected by covariates.

1.3 Thesis Outline

This thesis is organized as follows. In Chapter 2, the background methodology and the statistical tools used in the rest of the thesis are reviewed. The three main chapters illustrate our statistical proposals and they are arranged in order of publication. Chapter 3 discusses our novel longitudinal drift-diffusion mixed model for with an application to tone learning. Chapter 4 extends the methodological framework to longitudinal mixed models with locally informative predictors. Chapter 5 introduces a bivariate survival regression model for current status data. Concluding remarks and possible future research avenues are presented in Chapter 6.

For the sake of coherence, only three main methodological chapters were included in this thesis. However, as part of my doctoral research I have also worked on dependent partition models for single-cell RNA sequencing data (manuscript in preparation), posterior summaries of random partitions that favor repulsiveness (invited discussion of Wade and Ghahramani (2018)), flexible logistic mixed-effects models for learning curves (Paulon *et al.*, 2019), as well as the application of the drift-diffusion model on different learning modalities (Roark *et al.*, 2021).

Chapter 2

Background

In this chapter, we review the statistical methodologies upon which our contributions are based. Chapter 3 and Chapter 4 both use semiparametric regression techniques similar to the ones described in Section 2.2. Chapter 5 proposes a fully nonparametric survival regression based on the methods illustrated in Section 2.3. All chapters in this thesis use finite or infinite mixture models, whose main ideas are covered in Section 2.4. In order to accommodate time-varying partitions, Chapter 3 and Chapter 4 use (factorial) hidden Markov models (Section 2.5 and 2.6). Chapter 4 also uses ideas from tensor factorization methods (Section 2.7) to achieve significant reduction in model size.

2.1 Density Estimation via the Dirichlet Process

Density estimation problems involve inference about an unknown distribution H on the basis of an observed i.i.d. sample,

$$y_i \mid H \overset{\text{iid}}{\sim} H, \quad i = 1, \ldots, n.$$

If we wish to proceed with Bayesian inference, we need to complete the model with a prior probability model for the unknown "parameter" H. Assuming a prior model

on H requires the specification of a probability model for an infinite-dimensional parameter, that is, a BNP prior.

One of the most popular BNP models is the Dirichlet process (DP) prior. We refer to Müller *et al.* (2015, Chapter 2) for a full review of DP and DP mixtures as well as their practical applications. The DP was originally introduced by Ferguson (1973) and can be defined from its finite-dimensional analog. We write $H \sim DP(M, H_0)$ if the random distribution H is such that for any partition $A_1, \ldots, A_K$ of the sample space the random vector of the $H(A_i)$ follows a Dirichlet distribution, $(H(A_1), \ldots, H(A_K)) \sim \text{Dir}\{MH_0(A_1), \ldots, MH_0(A_K)\}$. The DP prior is indexed by the total mass parameter M (which controls the variance) and by the centering measure H_0 (which defines the expectation). In fact, $\text{E}[H(A)] = H_0(A)$ and $\text{Var}[H(A)] = H_0(A)\{1 - H_0(A)\}/(M + 1)$. Alternatively, Sethuraman (1994) gives a constructive definition for the DP, known as the stick-breaking construction:

$$H = \sum_{k=1}^{+\infty} \pi_k \delta_{\theta_k} \tag{2.1}$$

with $\pi_k = q_k \prod_{l<k}(1 - q_l), q_k \overset{\text{iid}}{\sim} \text{Beta}(1, M)$ and $\theta_k \overset{\text{iid}}{\sim} H_0$.

The DP generates almost surely discrete probability measures. For this reason, often an additional convolution with a continuous kernel $k(y \mid \theta)$ is used to represent a random probability measure

$$f_H(y) = \int k(y \mid \theta) dH(\theta) = \sum_{k=1}^{+\infty} \pi_k k(y \mid \theta_k)$$

with $H \sim DP(M, H_0)$. The model is known as DP mixture (DPM). The discrete nature of the DP draws can be useful when clustering of the observations is the main focus of the analysis. See Section 2.4 for more details.

2.2 Semiparametric Regression via B-splines

The generic regression framework attempts to explain an outcome variable y_i as a function of a covariate $x_i \in \mathcal{X}$. For ease of exposition we assume that both y_i and x_i are univariate. The regression problem can be stated as

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $f(\cdot)$ is an unknown mean function and $\varepsilon_i \sim f_\varepsilon$ are residuals. Usually, residuals are assumed to be independent and to satisfy $\mathbb{E}_{f_\varepsilon}(\varepsilon_i) = 0$. If the function f and the residual distribution are indexed by a finite dimensional parameter vector, then the problem reduces to traditional parametric regression, for example, normal linear regression. Bayesian nonparametric regression methods are available when relaxing restrictive parametric assumptions is of paramount importance.

A possible approach is to use flexible priors for random mean functions f, such as Gaussian process priors. However, many approaches do not define the probability model directly on f but instead expand the mean function on predetermined basis functions. Consider a function basis $\mathbf{B}(t) = \{B_1(t), \ldots, B_K(t)\}^\intercal$ and let $\boldsymbol{\beta}$ denote the set of coefficients with respect to that basis, i.e.,

$$f(x) = \sum_{k=1}^{K} \beta_k B_k(x)$$

Defining a probability model on $\boldsymbol{\beta}$ implicitly defines a probability model on the unknown mean function. Note that a fully nonparametric prior would require $K = \infty$. However, in practice only few of the weighted basis functions matter and the sum is truncated in some fashion.

A very flexible and popular basis expansion approach is based on B-splines (de Boor, 1978). A B-spline is a particular piecewise-differentiable polynomial of a given degree d, typically fixed at $d = 2$ or $d = 3$. B-splines are typically used with a rather large number K of basis functions. To avoid overfitting, a global level of smoothness is incorporated into such models by encouraging neighboring coefficients to be similar; the more regular the coefficients are, the less wiggly f is. A popular approach introduced in Eilers and Marx (1996) penalizes $\sum_{k=1}^{K}(\Delta^2\beta_k)^2$, the sum of squares of the second order differences in $\boldsymbol{\beta}$.

2.2.1 Linear B-splines

The construction of linear B-spline bases is detailed below. Consider knot points $t_1 = t_2 = A < t_3 < \cdots < B = t_{K+2} = t_{K+3}$ that divide $[A, B]$ into K equal subintervals, where $t_{2:(K+2)}$ are equidistant with $\delta = (t_3 - t_2)$. For $j = 2, 3, \ldots, K$, linear B-splines $B_{1,j}$ are then defined as

$$B_{1,j}(t) = \begin{cases} (t - t_j)/\delta & \text{if } t_j \le t < t_{j+1}, \\ (t_{j+2} - t)/\delta & \text{if } t_{j+1} \le t < t_{j+2}, \\ 0 & \text{otherwise.} \end{cases}$$

The components at the ends are likewise defined as

$$B_{1,1}(t) = \begin{cases} (t_3 - t)/\delta & \text{if } t_2 \le t < t_3, \\ 0 & \text{otherwise.} \end{cases}$$

$$B_{1,K+1}(t) = \begin{cases} (t - t_{K+2})/\delta & \text{if } t_{K+1} \le t < t_{K+2}, \\ 0 & \text{otherwise.} \end{cases}$$

Figure 2.1 shows the B-spline bases described here.

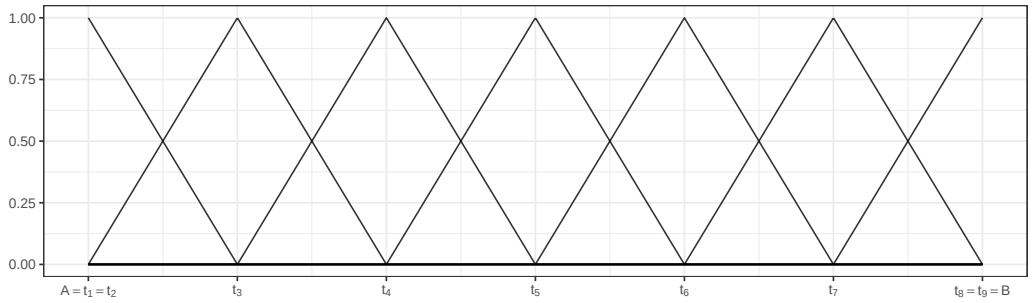


Figure 2.1: Plot of 7 linear B-splines on an interval $[A, B]$ defined by 9 equidistant knot points that divide $[A, B]$ into 6 equal subintervals.

2.2.2 Quadratic B-splines

The construction of quadratic B-spline bases is detailed below. Consider knot-points $t_1 = t_2 = t_3 = A < t_4 < \cdots < B = t_{K+3} = t_{K+4} = t_{K+5}$, where $t_{3:(K+3)}$ are equidistant with $\delta = (t_4 - t_3)$. For $j = 3, 4, \ldots, (K+2)$, quadratic B-splines $B_{2,j}$ are then defined as

$$B_{2,j}(X) = \begin{cases} \{(X - t_{J-1})/\delta\}^2/2 & \text{if } t_{J-1} \le X < t_J, \\ -\{(X - t_J)/\delta\}^2 + (X - t_J)/\delta + 1/2 & \text{if } t_J \le X < t_{j+2}, \\ \{1 - (X - t_{j+2})/\delta\}^2 & \text{if } t_{j+2} \le X < t_{j+3}, \\ 0 & \text{otherwise.} \end{cases}$$

The components at the ends are likewise defined as

$$B_{2,1}(X) = \begin{cases} \{1 - (X - t_1)/\delta\}^2/2 & \text{if } t_3 \le X < t_4, \\ 0 & \text{otherwise.} \end{cases}$$

$$B_{2,2}(X) = \begin{cases} -\{(X - t_3)/\delta\}^2 + (X - t_4)/\delta + 1/2 & \text{if } t_3 \le X < t_4, \\ \{1 - (X - t_4)/\delta\}^2/2 & \text{if } t_4 \le X < t_5, \\ 0 & \text{otherwise.} \end{cases}$$

$$B_{2,K+1}(X) = \begin{cases} \{(X - t_{K+1})/\delta\}^2/2 & \text{if } t_{K+1} \leq X < t_{K+2}, \\ -\{(X - t_{K+2})/\delta\}^2 + (X - t_{K+2})/\delta + 1/2 & \text{if } t_{K+2} \leq X < t_{K+3}, \\ 0 & \text{otherwise.} \end{cases}$$

$$B_{2,K+2}(X) = \begin{cases} \{(X - t_{K+2})/\delta\}^2/2 & \text{if } t_{K+2} \leq X < t_{K+3}, \\ 0 & \text{otherwise.} \end{cases}$$

Figure 2.2 provides a graphical illustration of these functions.

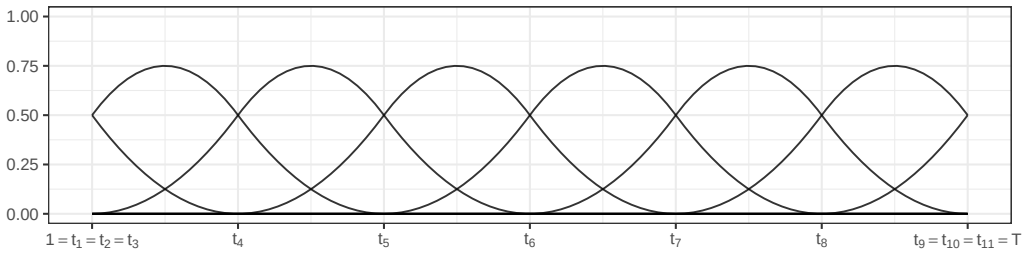


Figure 2.2: Plot of 8 quadratic B-splines on an interval $[1, T]$ defined by 11 knot points that divide $[1, T]$ into 6 equal subintervals.

2.3 Fully Nonparametric Regression

Fully nonparametric Bayesian regression methods allow both the mean function and the residual distribution to be modeled nonparametrically. These approaches, also known as density regression, imply that the complete shape of the response distribution is allowed to change as a function of the predictors. A general statement of the generic regression problem is

$$y_i \mid x_i \sim H_{x_i}$$

with a family $\mathcal{H} = \{H_x, x \in \mathcal{X}\}$ of probability measures indexed by $x \in \mathcal{X}$. The model is completed with a BNP prior on $\mathcal{H}$. A meaningful prior on $\mathcal{H}$ needs to

include dependence of the H_x's.

One of the most popular prior models for a family of random probability measures is the dependent DP (DDP), first introduced in MacEachern (1999). The stick-breaking representation (2.1) can be extended to DDP realizations as

$$H_x = \sum_{k=1}^{+\infty} \pi_k \delta_{\mu_{xk}}.$$

In the DDP model dependence is induced by assuming that, for fixed k, the atoms μ_{xk} are realizations of a Gaussian process, indexed by x. Independence across k, together with the stick-breaking prior for the common weights π_k (not indexed by x), maintains the marginal DP prior on H_x. This instance of the DDP model is known as "common weights DDP." Alternative implementations are possible, with dependent (across x) weights π_{xk} and common atoms ("common atoms DDP"), or the most general DDP model with dependent weights and atoms.

In the case of multiple categorical predictors $\mathbf{x}$, a simpler form of priors on family of random probability measures is a normal linear ANCOVA model for the atoms (De Iorio *et al.*, 2004), i.e. $\{\mu_{xk}, \mathbf{x} \in \mathcal{X}\}$. For instance, if $x_1 \in \{0, 1\}$ and $x_2 \in \{0, 1\}$, a simple model is

$$\mu_{\mathbf{x}k} = \delta_k + \alpha_k x_1 + \beta_k x_2,$$

In general, letting $\mathbf{d} = (1, x_1, x_2)^\intercal$ denote a design vector to select the desired ANOVA effects we can write $\mu_{\mathbf{x}_i k} = \mathbf{d}_i^\intercal \mathbf{m}_k$, where $\mathbf{m}_k = (\delta_k, \alpha_k, \beta_k)^\intercal$ are the linear model coefficients. This model induces the desired dependence of $H_{\mathbf{x}}$ across $\mathbf{x}$ by sharing, for example, the same β_k for any two covariate vectors $\mathbf{x}$ and $\mathbf{x}'$ that share the same x_2.

2.4 Clustering and Bayesian Mixture Models

A Bayesian mixture model assumes that data $\mathbf{y} = \{y_i\}_{i=1}^n$ are drawn from a density modeled as a convex combination of specified parametric families f, i.e.

$$f(y_i) = \sum_{k=1}^{K} \pi_k f(y_i \mid \theta_k).$$

Usually the component density f is fixed and known, whereas the component-specific parameters θ_k and the weights π_k are considered to be unknown. The number of components K is often assumed to be finite, in which case it is given a prior distribution. However, K can also be assumed to be infinity in the population. In this latter case, a stochastically decreasing prior on the weights is usually adopted and the finite sample will only use a finite number of components.

Mixture models may be adopted under a belief that the data are clustered, with observations within each cluster drawn from some simple parametric distribution. An alternative representation of Bayesian mixture models is the following. Suppose the population from which we are sampling is heterogeneous: there are multiple groups (or "clusters"), indexed by $k = 1, \ldots, K$, present in the population in proportions π_k, $k = 1, \ldots, K$. When sampling from group k, observations are assumed drawn from the density $f(\cdot \mid \theta_k)$. Equivalently,

$$y_i \mid z_i \sim f(\theta_{z_i}) \text{ with } p(z_i = k) = \pi_k.$$

The term "cluster" simply suggests both a degree of homogeneity within a cluster and a degree of separation between clusters. Hidden Markov models (see Section 2.5) can also be interpreted as finite mixture models with serial dependence in the

clustering allocations. We use HMM induced clustering in Chapter 3 and Chapter 4.

An important implication of the stick-breaking representation of DP draws (Section 2.1) is the fact that some values are coincident with positive probability. Thus, a partition of the indexes $\{1, 2, \ldots, n\}$, denoted with $\boldsymbol{\rho}_n = \{S_1, \ldots, S_K\}$, and therefore a clustering structure are induced by the ties in the sample, where K is the number of unique values in the sample and $S_j = \{i \in \{1, \ldots, n\} \text{ s.t. } Y_i = Y_j\}$ are the indexes of the j^{th} group. Let us denote with $\mathbf{n} = (n_1, \ldots, n_K)$ the cluster sizes for a partition of n observations into clusters $S_1, \ldots, S_K$. The prior distribution induced on $\boldsymbol{\rho}_n$ by $(Y_1, \ldots, Y_n)$ is in this case

$$p(\boldsymbol{\rho}_n) = \frac{\Gamma(M)}{\Gamma(M+n)} M^K \prod_{k=1}^{K} \Gamma(n_j),$$

where $\mathbf{n}$ is such that $\sum_{j=1}^{K} n_j = n$. This prior distribution on random partition is also known as exchangeable product partition function (EPPF). Chapter 5 uses random partitions induced by a DP prior model.

2.5 Hidden Markov Model (HMM)

The basic HMM (Frühwirth-Schnatter, 2006; McDonald and Zucchini, 1997, etc.) consists of two processes: an *observed* process $\{\mathbf{y}_t\}$ recorded sequentially over a set of discrete time points $t = 1, 2, \ldots, T$ and an associated *hidden* process $\{z_t\}$ which evolves according to a first order Markov chain with discrete state space. Specifically, an HMM makes the following set of conditional independence assumptions to model

the hidden and the observed processes

$$p(z_t \mid \mathbf{z}_{1:(t-1)}) = p(z_t \mid z_{t-1}),$$
$$p(\mathbf{y}_t \mid \mathbf{y}_{1:(t-1)}, \mathbf{z}_{1:t}) = p(y_t \mid z_t).$$

The distributions $p(z_t \mid z_{t-1})$ and $p(\mathbf{y}_t \mid z_t)$ are often referred to as the *transition distribution* and the *emission distribution*, respectively. Thus, the HMM admits a factorization of the joint distribution of $(\mathbf{y}_{1:T}, \mathbf{z}_{1:T})$:

$$p(\mathbf{y}_{1:T}, \mathbf{z}_{1:T}) = \pi_0(z_1)p(\mathbf{y}_1 \mid z_1) \prod_{t=2}^{T} p(z_t \mid z_{t-1})p(\mathbf{y}_t \mid z_t),$$

where π_0 denotes the distribution of the initial hidden variable z_1.

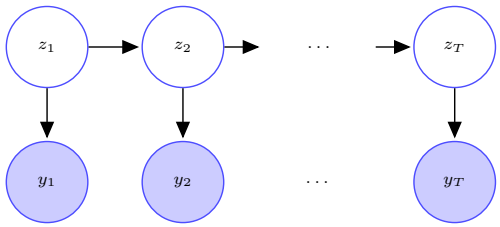


Figure 2.3: Directed acyclic graph (DAG) of an HMM.

Under a Bayesian set-up, efficient algorithms have been proposed in order to sample from the posterior distribution of the hidden states $\mathbf{z}_{1:T}$. As one can see in Figure 2.3, the joint conditional posterior distribution of the latent states can be factorized as

$$p(\mathbf{z}_{1:T} \mid \mathbf{y}_{1:T}, \boldsymbol{\zeta}) = p(z_T \mid z_{T-1}, \mathbf{y}_{1:T}, \boldsymbol{\zeta}) \ldots p(z_2 \mid z_1, \mathbf{y}_{1:T}, \boldsymbol{\zeta})p(z_1 \mid \mathbf{y}_{1:T}, \boldsymbol{\zeta}),$$

where $p(z_t \mid z_{t-1}, \mathbf{y}_{1:T}, \boldsymbol{\zeta}) \propto m_t(z_t)p(\mathbf{y}_t \mid z_t, \boldsymbol{\zeta})p(z_t \mid z_{t-1}, \boldsymbol{\zeta})$ and the backward messages are defined as $m_t(z_t) = p(\mathbf{y}_{t+1}, \mathbf{y}_{t+2}, \ldots, \mathbf{y}_T \mid z_t)$. To sample $\mathbf{z}_{1:T}$ from its

full conditional we first pass messages $m_t(z_t)$ backwards and then sample forwards. Recursion relations for backward messages can be derived using the conditional independence properties of the chain, i.e.

$$\begin{aligned} m_t(z_t) &= p(\mathbf{y}_{t+1}, \mathbf{y}_{t+2}, \ldots, \mathbf{y}_T \mid z_t) \\ &= \sum_{z_{t+1}} p(\mathbf{y}_{t+1}, \mathbf{y}_{t+2}, \ldots, \mathbf{y}_T, z_{t+1} \mid z_t) \\ &= \sum_{z_{t+1}} p(\mathbf{y}_{t+1} \mid z_{t+1}) p(\mathbf{y}_{t+2}, \ldots, \mathbf{y}_T \mid z_{t+1}) p(z_{t+1} \mid z_t) \\ &= \sum_{z_{t+1}} p(\mathbf{y}_{t+1} \mid z_{t+1}) p(z_{t+1} \mid z_t) m_{t+1}(z_{t+1}), \end{aligned}$$

with final condition

$$m_T(z_T) = 1.$$

Chapter 3 adapts such message passing algorithm to a different conditional independence structure.

2.6 Factorial HMM (fHMM)

In factorial HMMs (Ghahramani and Jordan, 1997), the latent states are represented by a collection of variables $\{\mathbf{z}_t\} = \{(z_t^{(1)}, \ldots, z_t^{(m)})\}$ where each component $\{z_t^{(\ell)}\}$ now evolves according to a first order Markov chain with discrete state spaces, and the *observed* process $\{y_t\}$ is observed sequentially as before over a set of discrete time points $t = 1, 2, \ldots, T$. An fHMM thus makes the following set of conditional independence assumptions to model the hidden and the observed processes

$$\begin{aligned} &p(\mathbf{z}_t \mid \mathbf{z}_{1:(t-1)}) = \textstyle\prod_{\ell=1}^m p(z_t^{(\ell)} \mid z_{t-1}^{(\ell)}), \\ &p(y_t \mid \mathbf{y}_{1:(t-1)}, \mathbf{z}_{1:t}) = p(y_t \mid \mathbf{z}_t) = p(y_t \mid z_t^{(1)}, \ldots, z_t^{(m)}). \end{aligned}$$

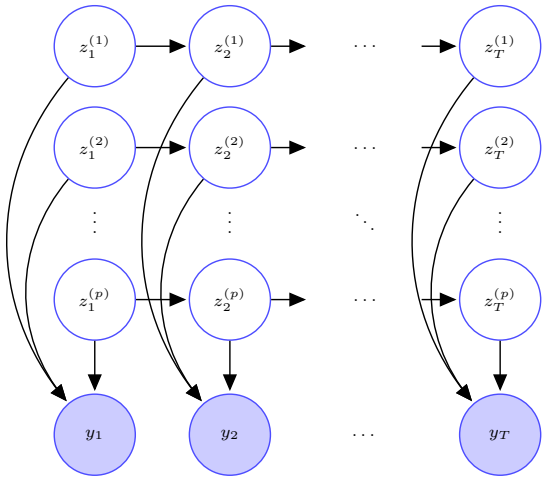


Figure 2.4: Directed acyclic graph (DAG) of a fHMM with p layers.

In Chapter 3, we adapt the basic fHMM to characterize local influences of categorical predictors in longitudinal functional models. For each categorical predictor $x_j \in \{1, \ldots, x_{j,\max}\}$, we introduce an fHMM $\{\mathbf{z}_{j,t} = (z_{j,t}^{(1)}, \ldots, z_{j,t}^{(x_{j,\max})})\}$ with $x_{j,\max}$ layers, one for each level of x_j. Conditional on $(z_{1,t}^{(x_1)}, \ldots, z_{p,t}^{(x_p)}) = (z_{1,t}, \ldots, z_{p,t})$, we then associate the coefficients $\beta_{t,x_1,\ldots,x_p}$ of a predictor dependent B-spline mixture model with atoms $\beta^{\star}_{t,z_{1,t},\ldots,z_{p,t}}$. Specifically, we let

$$p(\mathbf{z}_t \mid \mathbf{z}_{1:(t-1)}) = \textstyle\prod_{j=1}^{p} p(\mathbf{z}_{j,t} \mid \mathbf{z}_{j,t-1}) = \prod_{j=1}^{p} \prod_{\ell=1}^{m} p(z_{j,t}^{(\ell)} \mid z_{j,t-1}^{(\ell)}),$$
$$\{\beta_{t,x_1,\ldots,x_p} \mid z_{j,t}^{(x_j)} = z_{j,t}, j = 1, \ldots, p\} = \beta^{\star}_{t,z_{1,t},\ldots,z_{p,t}}.$$

Forward-backward (or backward-forward) algorithms for HMMs rely on passing messages forward (or backward) and then sampling backward (or forward) (Rabiner, 1989; Scott, 2002). While adapting such algorithms to fHMMs, the requirement to sum over all possible configurations in computing the messages becomes a challenge. Hamming ball samplers for fHMMs (Titsias and Yau, 2014) avoid

this computationally expensive step by introducing and conditioning on an auxiliary variable that restricts the sampling to only a slice (Neal, 2003) of the entire high-dimensional space. In doing so, the sampler also allows localized joint updating of all constituent chains, making it less prone to get trapped in local modes.

Let $h(\mathbf{z}_t, \mathbf{v}_t) = \sum_{\ell=1}^{L} 1\{z_t^\ell \neq v_t^\ell\}$ denote the Hamming distance between the vectors $\mathbf{z}_t = (z_t^{(1)}, \ldots, z_t^{(L)})^{\mathrm{T}}$ and $\mathbf{v}_t = (v_t^{(1)}, \ldots, v_t^{(L)})^{\mathrm{T}}$, and let $\mathcal{H}_m(\mathbf{z}_t) = \{\mathbf{v}_t : h(\mathbf{z}_t, \mathbf{v}_t) \leq m\}$ denote a Hamming ball of radius m around $\mathbf{z}_t$. Consider an fHHM, as shown in Figure 2.4 but with L component chains each with state space $\{1, \ldots, d\}$. Introducing an auxiliary variable $\mathbf{v}$ following a conditional probability distribution $p(\mathbf{v} \mid \mathbf{z}) = \prod_{t=1}^{T} p(\mathbf{v}_t \mid \mathbf{z}_t)$, the augmented joint model becomes $p(\mathbf{y}, \mathbf{z}, \mathbf{v}) = p(\mathbf{v} \mid \mathbf{z})p(\mathbf{y} \mid \mathbf{z})p(\mathbf{z}) = \big\{ \prod_{t=1}^{T} p(\mathbf{v}_t \mid \mathbf{z}_t)p(\mathbf{y}_t \mid \mathbf{z}_t) \big\} p(\mathbf{z}_1) \prod_{t=2}^{T} p(\mathbf{z}_t \mid \mathbf{z}_{t-1})$. Sampling $\mathbf{v}$ from the posterior can then be done by sampling independently from the full conditionals $p(\mathbf{v}_t \mid \mathbf{z}_t)$. Sampling $\mathbf{z}$ from the posterior can still be carried out using forward-backward (or backward-forward) message passing algorithms but with the augmented full conditional $p(\mathbf{z} \mid \mathbf{y}, \mathbf{v}) \propto \big\{ \prod_{t=1}^{T} p(\mathbf{v}_t \mid \mathbf{z}_t)p(\mathbf{y}_t \mid \mathbf{z}_t) \big\} \big\{ \prod_{t=2}^{T} p(\mathbf{z}_t \mid \mathbf{z}_{t-1}) \big\} p(\mathbf{z}_1)$.

The set of possible configurations needed to compute the messages at time t is now restricted to the support of $p(\mathbf{v}_t \mid \mathbf{z}_t)$. If this can be made much smaller compared to the original size of the state space, computational burden can be greatly reduced. The Hamming ball algorithm does this by setting $p(\mathbf{v}_t \mid \mathbf{z}_t) \propto 1\{\mathbf{v}_t \in \mathcal{H}_m(\mathbf{z}_t)\}$, that is, by sampling the $\mathbf{v}_t$'s uniformly from $\mathcal{H}_m(\mathbf{z}_t)$. By symmetry, since $\mathbf{v}_t \in \mathcal{H}_m(\mathbf{z}_t)$ if and only if $\mathbf{z}_t \in \mathcal{H}_m(\mathbf{v}_t)$, the support of each $\mathbf{z}_t$ in the full conditional $p(\mathbf{z} \mid \mathbf{y}, \mathbf{v})$ is then restricted only to $\mathcal{H}_m(\mathbf{v}_t)$.

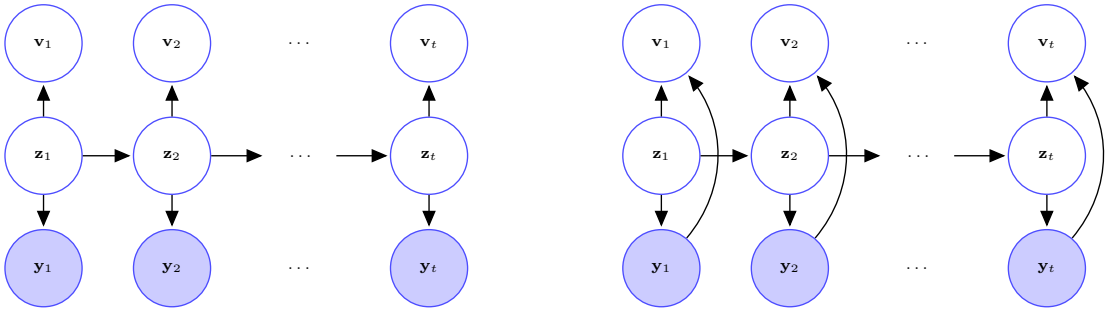


Figure 2.5: Graph of a Hamming ball sampler (left panel) and a locally informed Hamming ball sampler (right panel) for fHMM.

The Hamming ball sampler is still limited in its ability to efficiently explore the neighborhood of $\mathbf{z}_t$ as it blindly proposes new values along arbitrarily chosen directions within the ball. More informed moves can be proposed utilizing the information contained in the likelihood function (Zanella, 2019). For instance, $p(\mathbf{v}_t \mid \mathbf{z}_t, \mathbf{y}_t) \propto g\{p(\mathbf{y}_t \mid \mathbf{v}_t)\}1\{\mathbf{v}_t \in \mathcal{H}_m(\mathbf{z}_t)\}$, for proper choices of $g(\cdot)$, favors moves along directions that increase the conditional likelihood $p(\mathbf{y}_t \mid \mathbf{v}_t)$ (Figure 2.5). The augmented joint model now becomes $p(\mathbf{y}, \mathbf{z}, \mathbf{v}) = p(\mathbf{v} \mid \mathbf{y}, \mathbf{z})p(\mathbf{y} \mid \mathbf{z})p(\mathbf{z}) = \big\{\prod_{t=1}^{T} p(\mathbf{v}_t \mid \mathbf{y}_t, \mathbf{z}_t)p(\mathbf{y}_t \mid \mathbf{z}_t)\big\}\big\{\prod_{t=2}^{T} p(\mathbf{z}_t \mid \mathbf{z}_{t-1})\big\}p(\mathbf{z}_1)$. Sampling $\mathbf{z}$ from the posterior can be carried out using message passing algorithms as before with each $\mathbf{z}_t$ restricted to $\mathcal{H}_m(\mathbf{v}_t)$ but with the updated full conditionals $p(\mathbf{z} \mid \mathbf{y}, \mathbf{v}) \propto \big\{\prod_{t=1}^{T} p(\mathbf{v}_t \mid \mathbf{z}_t, \mathbf{y}_t)p(\mathbf{y}_t \mid \mathbf{z}_t)\big\}\big\{\prod_{t=2}^{T} p(\mathbf{z}_t \mid \mathbf{z}_{t-1})\big\}p(\mathbf{z}_1)$.

2.7 Tensor Factorization Methods

In this section, we provide a brief review of the different main types of tensor factorizations (Hitchcock, 1927; Tucker, 1966; De Lathauwer *et al.*, 2000; Kolda and

Bader, 2009).

A $d_1 \times \cdots \times d_p$ dimensional tensor $\boldsymbol{\beta} = \{\beta_{h_1,\ldots,h_p} : h_j = 1, \ldots, d_j, j = 1, \ldots, p\}$ admits a parallel factor (PARAFAC) decomposition with rank r (Figure 2.6) if it can be written as

$$\beta_{h_1,\ldots,h_p} = \sum_{z=1}^{r} \prod_{j=1}^{p} a_{j,z}^{(h_j)} \quad \text{for each } (h_1, \ldots, h_p), \tag{2.2}$$

where $\mathbf{a}_{j,z} = \{a_{j,z}^{(h_j)} : h_j = 1, \ldots, d_j\}, z = 1, \ldots, r, j = 1, \ldots, p$ are $d_j \times 1$ dimensional vectors.

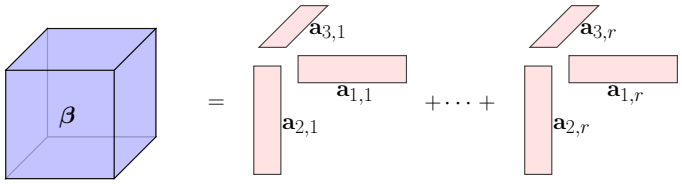


Figure 2.6: Pictorial representation of PARAFAC of a three dimensional tensor.

A $d_1 \times \cdots \times d_p$ dimensional tensor $\boldsymbol{\beta} = \{\beta_{h_1,\ldots,h_p} : h_j = 1, \ldots, d_j, j = 1, \ldots, p\}$ admits a Tucker decomposition with multi-linear rank $(r_1, \ldots, r_p)$ (Figure 2.7) if it admits a representation

$$\beta_{h_1,\ldots,h_p} = \sum_{z_1=1}^{r_1} \cdots \sum_{z_p=1}^{r_p} \beta^{\star}_{z_1,\ldots,z_p} \prod_{j=1}^{p} a_{j,z_j}^{(h_j)} \quad \text{for each } (h_1, \ldots, h_p), \tag{2.3}$$

where $\boldsymbol{\beta}^{\star} = \{\beta^{\star}_{z_1,\ldots,z_p} : z_j = 1, \ldots, r_j, j = 1, \ldots, p\}$ is an $r_1 \times \cdots \times r_p$ dimensional 'core tensor' with $1 \leq r_j \leq d_j$ for each j, and $\mathbf{A}_j = \{a_{j,z_j}^{(h_j)} : h_j = 1, \ldots, d_j, z_j = 1, \ldots, r_j\}, j = 1, \ldots, p$ are $d_j \times r_j$ dimensional 'mode matrices' or 'factor matrices' with full column rank r_j. The effective size of the model after the factorization is $\prod_{j=1}^{p} r_j + \sum_{j=1}^{p} r_j d_j \approx \prod_{j=1}^{p} r_j$. A significant reduction in dimensions is therefore

achieved by the decomposition when $\prod_{j=1}^{p} r_j \ll \prod_{j=1}^{p} d_j$, that is, the size of the core tensor is much smaller than the size of the original tensor.

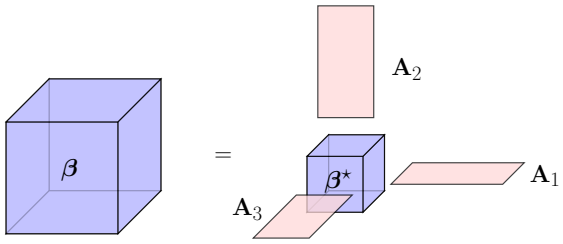


Figure 2.7: Pictorial representation of HOSVD of a three dimensional tensor.

The PARAFAC representation is obtained as a special case of the Tucker decomposition with $r_1 = \cdots = r_p = r$ and $\beta_{h_1,\ldots,h_p} = 1\{h_1 = \cdots = h_p\}$. Compared to the PARAFAC, the Tucker decomposition thus typically achieves a much greater reduction in the dimension of a tensor.

The compact higher order singular value decomposition (compact HOSVD) of a tensor is a special case of the Tucker decomposition, where the mode matrices $\mathbf{A}_j$'s are restricted to be semi-orthogonal, that is, they satisfy $\mathbf{A}_j^{\mathrm{T}}\mathbf{A}_j = \mathbf{I}_{r_j}$ for all j.

While none of these representations are fully identifiable, the compact HOSVD results in an equally flexible but much more interpretable form of the Tucker decomposition.

In formulating our model for the fixed effects in Chapter 4, we structure the parameters for different predictor combinations as a $x_{1,\max} \times \cdots \times x_{p,\max}$ dimensional tensor $\boldsymbol{\beta}_k = \{\beta_{k,x_1,\ldots,x_p} : (x_1,\ldots,x_p) \in \mathcal{X}\}$ for different predictor combinations at each location k and then apply a compact HOSVD-type factorization (Figure 4.4)

as

$$\{\beta_{k,x_1,\ldots,x_p} \mid z_{j,k}^{(x_j)}, j = 1,\ldots,p\} = \sum_{z_{1,k}} \cdots \sum_{z_{p,k}} \beta^{\star}_{k,z_{1,k},\ldots,z_{p,k}} \prod_{j=1}^{p} 1\{z_{j,k}^{(x_j)} = z_{j,k}\},$$

where $\boldsymbol{\beta}^{\star}_k = \{\beta^{\star}_{k,z_{1,k},\ldots,z_{p,k}} : (z_{1,k},\ldots,z_{p,k}) \in \mathcal{Z}_k\}$ is a $\ell_{1,k} \times \cdots \times \ell_{p,k}$ dimensional core tensor and $\mathbf{z}_{j,k} = \{1\{z_{j,k}^{(x_j)} = z_{j,k}\} : x_j \in \mathcal{X}_j, z_{j,k} \in \mathcal{Z}_{j,k}\}$ are $x_{j,\max} \times \ell_{j,k}$ dimensional allocation matrices with binary entries. This is a compact HOSVD-type factorization since $\mathbf{z}_{j,k}^{\mathrm{T}}\mathbf{z}_{j,k}$ are diagonal matrices for all $j = 1,\ldots,p, k = 1,\ldots,K$.

Chapter 3

Bayesian Semiparametric Longitudinal Drift-Diffusion Mixed Models for Tone Learning in Adults

This chapter is based on the publication Paulon *et al.* (2020) and it introduces a novel methodology for multi-alternative decision making in longitudinal settings. Understanding how adult humans learn non-native speech categories such as tone information has shed novel insights into the mechanisms underlying experience-dependent brain plasticity. Scientists have traditionally examined these questions via longitudinal learning experiments under a multi-category decision making paradigm. Drift-diffusion processes are popular in such contexts for their ability to mimic underlying neural mechanisms. Motivated by these problems, we develop a novel Bayesian semiparametric inverse Gaussian drift-diffusion mixed model for multi-alternative decision making in longitudinal settings. We design a Markov chain Monte Carlo algorithm for posterior computation. We evaluate the method's empirical performances through synthetic experiments. Applied to our motivating longitudinal tone learning study, the method provides novel insights into how the biologically inter-

---

This chapter appears as: *G. Paulon, F. Llanos, B. Chandrasekaran, A. Sarkar. Bayesian semiparametric longitudinal drift-diffusion mixed models for tone learning in adults. Journal of the American Statistical Association, 2020.* My contributions include data preparation and analysis, methods development, manuscript writing and preparation, and software implementation.

pretable model parameters evolve with learning, differ between input-response tone combinations, and differ between well and poorly performing adults.

3.1 Introduction

Understanding the cognitive and biological mechanisms underlying our ability to learn new speech categories in adulthood constitute important questions in auditory neuroscience. Recent studies have demonstrated that adults are capable of learning features of a second language to a high degree of efficiency, demonstrating that age need not always constrain language learning abilities. The inherent dynamic complexities underlying learning in adulthood are not yet well understood but are being studied through extensive ongoing research.

The research reported here is motivated particularly by experiments on the acquisition of Mandarin tones by native speakers of English. Native speech categories are acquired during the first year of life, within a so-called phonetic sensitivity period. There is a greater neural commitment to native-language speech sounds, and this commitment may preclude the learning of novel speech categories in adulthood (Johnson and Newport, 1989; Iverson *et al.*, 2003). In Mandarin Chinese, there are four tone categories that systematically change word meaning, similar to consonants and vowels in English. These tones are, however, linguistically irrelevant in English. English native speakers thus struggle to distinguish the four tones and generalize their differences (Wang *et al.*, 1999; Chandrasekaran *et al.*, 2010; Maddox and Chandrasekaran, 2014). In laboratory settings, combining exposure to perceptually variable tones with trial-by-trial corrective feedback can improve tone categorization

skills within a few hundred trials. Reaching a native like proficiency, however, may take several sessions of training (Xie *et al.*, 2017; Reetzke *et al.*, 2018). The perceptual and sensory representation of Mandarin tones gets fundamentally refined over the course of this learning period (Feng *et al.*, 2019). Understanding this longitudinal evolution is critical to assess the cognitive dynamics of speech category learning. The statistical challenge is to make this assessment indirectly from behavioral data on tone categorization responses and response times.

To this end, we identify the Mandarin tone categorization problem with the broader class of problems of multi-category decision making under perceptual stimuli (Smith and Ratcliff, 2004; Heekeren *et al.*, 2004; Gold and Shadlen, 2007; Schall, 2001; Purcell, 2013; Glimcher and Fehr, 2013). In such contexts, drift-diffusion processes are popular models for behavioral accuracies and response times as they mimic the accumulation of sensory evidence in favor of different decision alternatives in the human brain (Ratcliff, 1978; Ratcliff *et al.*, 2016). The existing literature on drift-diffusion models is substantive (Smith and Vickers, 1988; Ratcliff and Rouder, 1998; Ratcliff and McKoon, 2008). These classical methods, as well as their recent adaptations using reinforcement learning based ideas (Fontanesi *et al.*, 2019; Pedersen *et al.*, 2017; Peters and D'Esposito, 2020), are, however, heavily focused on the two category case with a single latent diffusion process and two boundaries, one for each of the two decision alternatives. This is despite the fact that humans often are required to learn more than two categories at once. For example, English has 14 vowels and 24 consonant phonemes; Mandarin has four tone categories, etc. The joint likelihood of accuracies and response times under models with a single diffusion process

is mathematically complex and computationally expensive (Navarro and Fuss, 2009; Tuerlinckx, 2004; Tuerlinckx *et al.*, 2001). Inference in such models is thus often based on approximations of the likelihood (Vandekerckhove and Tuerlinckx, 2007), or on the conditional likelihood of the response times, conditioned on the decisions (Vandekerckhove *et al.*, 2008). Multi-category drift-diffusion models with separate latent processes, one for each decision category and simultaneously at play, have been developed to address some of the limitations (Usher and McClelland, 2001; Brown and Heathcote, 2008; Leite and Ratcliff, 2010; Dufau *et al.*, 2012; Kim *et al.*, 2017), but the relevant literature remains sparse and focused only on simple static designs.

Learning to distinguish Mandarin tones or, more generally, to make categorization decisions is, however, a dynamic process, driven by continuous and nuanced perceptual adjustments in our brain and behavior over time. The existing simple static models are thus severely limited in their ability to capture the true inherent complexities, including assessing the biologically relevant changes that take place over the learning period. Principled statistical approaches to multi-category dynamic drift-diffusion mixed effects models, that appropriately accommodate fixed effects of experimental factors as well as random effects due to subjects, are therefore highly needed but present daunting methodological and computational challenges.

In this chapter, we address these challenges by developing a novel biologically interpretable flexible Bayesian semiparametric inverse Gaussian drift-diffusion mixed model for studying multi-alternative perceptual decision making processes in longitudinal settings.

Our construction proceeds by characterizing the accumulation of evidence for

different input-response tone combinations by associated independent Wiener diffusion processes, resulting in an inverse Gaussian distribution based joint probability model for the final response tone and the associated response time. To adapt this to a longitudinal mixed model setting, we then assume the model parameters to comprise input-response tone specific fixed effects and subject specific random effects, modeling them both by mixtures of locally supported B-spline bases (de Boor, 1978; Eilers and Marx, 1996) spanning the length of the longitudinal experiment. Both these effects are thus allowed to evolve flexibly as smooth functions over the training period (Ramsay and Silverman, 2007; Morris, 2015; Wang *et al.*, 2016) as the participants get more experience and training in their assigned decision tasks.

Dependence in the fixed effects model spline coefficients across adjacent temporal regions is induced via hidden Markov models (HMMs) (McDonald and Zucchini, 1997; Rabiner, 1989; Frühwirth-Schnatter, 2006; Cappé *et al.*, 2005), one for each input-response tone combination but all sharing a common state space, as well as a novel smoothness inducing Markovian prior on the core spline coefficients. The HMMs, adapted in such novel ways, induce a local clustering of the fixed effects spline coefficients associated with different input-response tone combinations, in effect, allowing us to assess local similarities and differences between the corresponding parameter trajectories in different learning phases.

This ability to infer local similarities and differences in the cognitive dynamics is theoretically and practically relevant for tone learning applications. The underlying mechanisms are expected to be very similar when the participants are first introduced to the tones; differences may appear as they get better at identifying the tones as

some tones may be easier to identify than others in this stage; these differences may start to disappear again in later stages of the experiment as the participants become highly proficient in identifying all the different tones. As for individual heterogeneity, neural measures of sensory encoding information collected prior to the learning task show no clear individual differences, even though the process of learning itself results in good and poor learners (Reetzke *et al.*, 2018).

The literature on longitudinal data analysis models is enormous. See, for example, books by Diggle *et al.* (2002); Singer *et al.* (2003); Fitzmaurice *et al.* (2008) and the references therein. Bayesian methods for longitudinal data have also been extensively developed (Daniels and Pourahmadi, 2002; Chib and Hamilton, 2002; Li *et al.*, 2010; Müller *et al.*, 2013; Quintana *et al.*, 2016, etc.). The problem of modeling locally clustered effects has, however, not garnered much attention. We can only mention Petrone *et al.* (2009); Nguyen and Gelfand (2011, 2014), all of which were designed primarily for normally distributed functional data with continuous covariates. It is not clear how these approaches can be adapted to our problem.

Overall, our proposed method takes the existing state-of-the-art many significant steps forward, including (a) introducing a novel biologically interpretable class of multi-category inverse Gaussian drift-diffusion models for decision making, (b) accommodating fixed effects of perceptual stimuli and random effects due to subject specific heterogeneity in such models in a statistically principled manner, (c) adapting these models to longitudinal study designs, studying the temporal evolution of the underlying process parameters as the subjects get trained and experienced in their assigned decision tasks, (d) allowing the process parameters to be locally clus-

tered, enabling the assessment of their similarities and differences in various learning stages.

Applied to our motivating tone learning data set, the proposed method provides many novel insights into the cognitive dynamics, allowing us to answer important scientific questions completely outside the scope of the previously existing literature. These include a detailed understanding of how biologically significant model parameters, that systematically relate to the underlying neural processes, evolve and interplay to enable gradual longitudinal learning in the participants, how similar or different these parameters are across different input and output tone combinations in different learning phases, how these processes differ between a good and a bad learner, etc.

Outline of the Chapter: The rest of this chapter is organized as follows. Section 3.2 introduces the real data and the relevant scientific background. Section 3.3 develops the generic longitudinal drift-diffusion mixed model. Section 3.4 develops a Markov chain Monte Carlo (MCMC) algorithm for posterior computation. Section 3.5 presents the real data application. Section 3.6 and 3.7 show comparisons with competing models. Section 3.8 presents the results of simulation studies. Section 3.9 contains concluding remarks.

3.2 Behavioral Data and Scientific Background

The behavioral data set that motivated our research comes from an intensive multi-day longitudinal speech category training study reported previously in Reetzke *et al.* (2018). In this study, $n = 20$ native English-speaking adults were trained to

categorize Mandarin Chinese syllables into lexical tone categories as a function of their pitch contour. Mandarin Chinese has four syllabic pitch contours or tones that are used to convey different lexical meanings. For example, in Mandarin Chinese, the syllable 'ma' can be interpreted as 'mother', 'hemp', 'horse', or 'scold' depending on whether is pronounced with a high-level (T1), low-rising (T2), low-dipping (T3), or high-falling (T4) tone, respectively. The stimuli consisted of these tones pronounced by four native Mandarin speakers. The trials were administered in homogeneous blocks. Each block comprised 40 categorization trials for 40 different speech exemplars, corresponding to different combinations of speakers, syllables, and input tones. Participants were trained across several days, with five blocks on each day. On each categorization trial, participants indicated the tone category they heard via a button press on a computer keyboard. Following the button press, the participants were given corrective feedback ('Correct/Incorrect') on a computer screen which was previously shown to be more effective in enhancing learning compared to full feedback (for example, 'Incorrect, that was a category 2') (Chandrasekaran *et al.*, 2014). Individual categorization performance was monitored across training sessions until each participant achieved and maintained accuracy levels comparable to that of native speakers of Mandarin.

The data consist of the tone responses and the associated response times for different input tones for the 20 participants. We focus here on the first two days of training (10 blocks in total) as they exhibited the steepest improvement in learning as well as the most striking individual differences relative to any other collection of blocks (Figure 3.1). In that sense, they provide an optimal longitudinal frame to

assess the effects of learning on decision making variables.

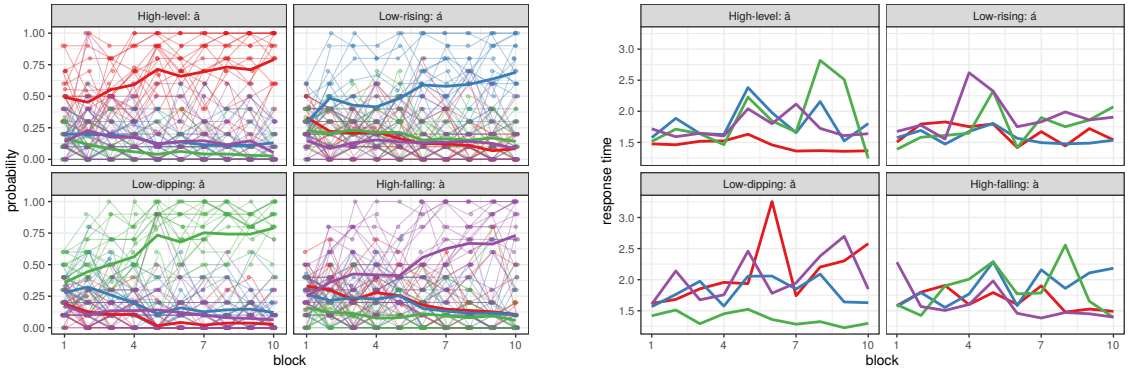


Figure 3.1: Left panel: Proportions of times an input tone was classified into different tone categories by different subjects. The thick line represents the average performance across subjects. Right panel: Associated response times averaged across subjects for clarity. In both panels, high-level tone responses are shown in red; low-rising in blue; low-dipping in green; and high-falling in purple.

Tone learning can be viewed from a broader perspective of multi-category decision making tasks, and hence can be studied using computational models developed for such tasks. We present here a brief nontechnical overview of how these models relate to the underlying neurobiology. Mathematical details and developments are deferred to Section 3.3.

In a typical multi-category decision task, the brain accumulates sensory evidence in order to make a categorical decision. This accumulation process is reflected in increasing firing rate at local neural populations associated with alternative decisions. A decision is taken when neural activity in one of these populations crosses a particular threshold level. The decision category that is finally chosen is the one whose decision threshold is crossed first (Gold and Shadlen, 2007; Brody and Hanks,

2016).

Changes in evidence accumulation rates and decision thresholds can be induced by task difficulty, neurostimulation, and/or individual differences in cognitive function (Cavanagh *et al.*, 2011; Ding and Gold, 2013). Decision-making is also regulated by demands on both speed and accuracy as a function of the task (Bogacz *et al.*, 2010; Milosavljevic *et al.*, 2010). The overall learning accuracies ('Correct/Incorrect' response proportions) in our data set were previously analyzed in Paulon *et al.* (2019) using a binary logistic longitudinal mixed model. In a different context, Craigmile *et al.* (2010) had developed a model for response times. Separate models for accuracies and response times cannot, however, provide a meaningful interpretation of the speed-accuracy trade-off.

An excellent basis for jointly modeling accuracies and response times is obtained by imitating the underlying neural evidence accumulation mechanisms via latent drift-diffusion processes racing toward their respective boundaries, the process reaching its boundary first producing the final observed decision and the time taken to reach this boundary giving the associated response time (Figure 3.2) (Usher and McClelland, 2001). The drift and the boundary parameters jointly explain the dynamics of choice, including the speed-accuracy trade-off. Broadly speaking, decision thresholds remaining fixed, higher drift rates lead to faster and more accurate responses; for fixed drift rates, higher decision thresholds, on the other hand, increase response times as well as inaccuracies.

In our motivating tone learning experiment, we are interested in understanding the evolution and interplay of the drift and the boundary parameters behind the

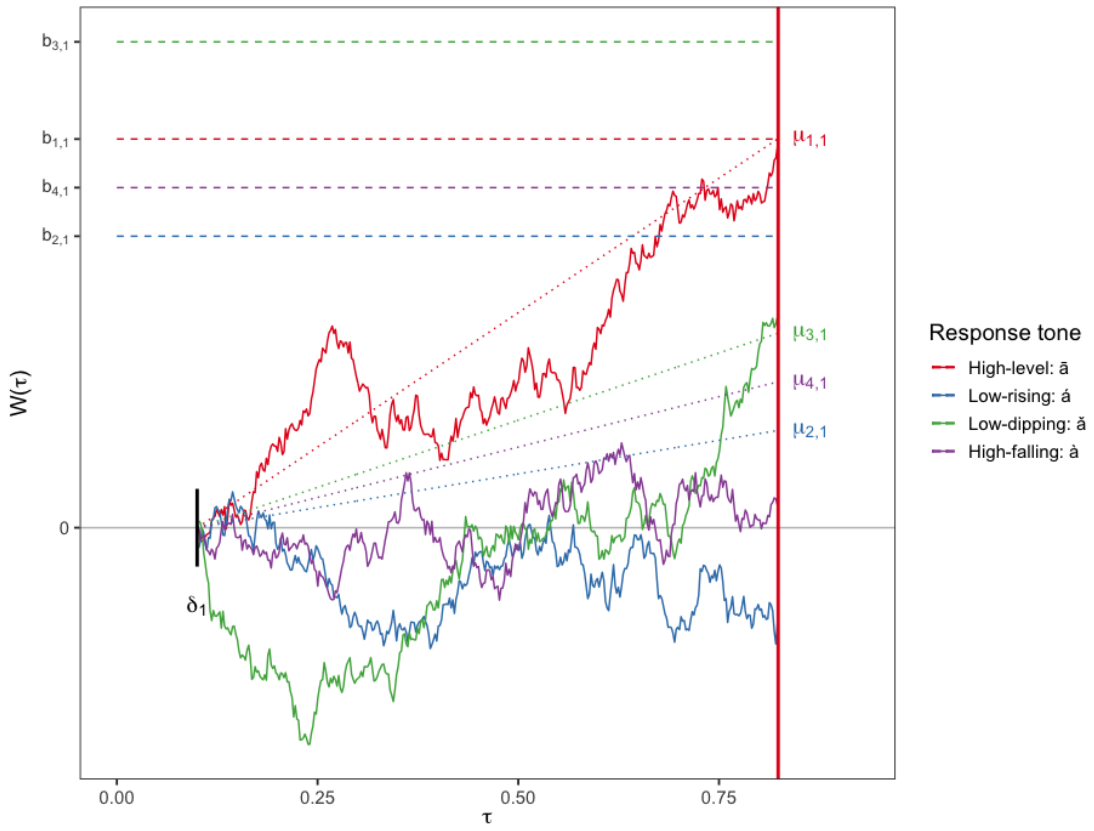


Figure 3.2: Drift-diffusion model for perceptual decision making. After an initial δ_s amount of time required to encode an input signal s, the evidence in favor of a response category d accumulates according to a Wiener diffusion process with drift $\mu_{d,s}$. The decision d is eventually taken if the underlying process is the first to reach its decision boundary $b_{d,s}$. Here we illustrate a tone learning trial with input tone T1 ($s = 1$) that was eventually correctly identified. Section 3.2 provides additional neurobiological background. Section 3.3 provides additional mathematical details.

improved tone identification performances over training. Importantly, as was also discussed in the introduction, we are not just interested in estimating the overall trajectories of these parameters but also how they might differ between different

input-response tone combinations locally in different longitudinal stages of the experiment. Additional interest lies in assessing subject level heterogeneity in these parameter trajectories, including particularly how they differ between good versus bad learners.

3.3 Longitudinal Drift-Diffusion Mixed Models

The basic Wiener diffusion process can be specified as $W(\tau) = \mu\tau + \sigma B(\tau)$, where $B(\tau)$ is the standard Brownian motion, μ is the drift rate, and σ is the diffusion coefficient (Cox and Miller, 1965; Ross *et al.*, 1996). The process has independent normally distributed increments, that is, $\Delta W(\tau) = \{W(\tau + \Delta\tau) - W(\tau)\} \sim \text{Normal}(\mu\Delta\tau, \sigma^2\Delta\tau)$, independently from $W(\tau)$. The first passage time of crossing a threshold b, $\tau = \inf\{\tau' : W(0) = 0, W(\tau') \geq b\}$, is then distributed according to an inverse Gaussian distribution (Whitmore and Seshadri, 1987; Chhikara, 1988; Lu, 1995) with density

$$f(\tau \mid \mu, \sigma^2, b) = \frac{b}{\sqrt{2\pi\sigma^2}}\tau^{-3/2}\exp\left\{-\frac{(b-\mu\tau)^2}{2\sigma^2\tau}\right\}, \quad b > 0, \quad \mu > 0, \quad \sigma^2 > 0.$$

With $\boldsymbol{\theta} = (\mu, \sigma, b)^{\mathrm{T}}$, we have $\mathbb{E}(\tau \mid \boldsymbol{\theta}) = b/\mu$ and $\text{var}(\tau \mid \boldsymbol{\theta}) = b\sigma^2/\mu^3$.

Given perceptual stimuli and a set of decision choices, the neurons in the brain accumulate evidence in favor of the different alternatives. Modeling this behavior using Wiener processes with unit variances, assuming that a response is given when the decision threshold for one of the options is crossed, a probability model for the time τ_d to reach the threshold for the d^{th} decision category under the influence of the

s^{th} stimulus is obtained as

$$f(\tau_d \mid \delta_s, \mu_{d,s}, 1, b_{d,s}) = \frac{b_{d,s}}{\sqrt{2\pi}}(\tau_d - \delta_s)^{-3/2} \exp\left[-\frac{\{b_{d,s}-\mu_{d,s}(\tau_d-\delta_s)\}^2}{2(\tau_d-\delta_s)}\right], \tag{3.1}$$

where $\mu_{d,s}$ denotes the rate of accumulation of evidence, $b_{d,s}$ the decision boundaries, and δ_s an offset representing the collective time required to encode the s^{th} signal before evidence accumulation begins, the time to press a computer key to record a response after a decision is reached, etc. (Figure 3.2). We now let $\boldsymbol{\theta}_{d,s} = (\delta_s, \mu_{d,s}, b_{d,s})^{\mathrm{T}}$. Since a decision d is reached at response time τ if the corresponding threshold is crossed first, that is when $\{\tau = \tau_d\} \cap_{d' \neq d} \{\tau_{d'} > \tau_d\}$, we have $d = \arg\min \tau_{d'}$. Assuming simultaneous accumulation of evidence for all decision categories, modeled by independent Wiener processes, and termination when the threshold for the observed decision category d is reached, the joint distribution of (d, τ) is thus given by

$$f(d, \tau \mid s, \boldsymbol{\theta}) = g(\tau \mid \boldsymbol{\theta}_{d,s}) \prod_{d' \neq d} \{1 - G(\tau \mid \boldsymbol{\theta}_{d',s})\}. \tag{3.2}$$

where, to distinguish from the generic notation f, we now use $g(\cdot \mid \boldsymbol{\theta})$ and $G(\cdot \mid \boldsymbol{\theta})$ to denote, respectively, the probability density function (pdf) and the cumulative distribution function (cdf) of an inverse Gaussian distribution, as defined in (3.1). We refer to model (3.2) as the inverse Gaussian drift-diffusion model.

The marginal distribution of the response times τ under the influence of stimulus s is then obtained as

$$f(\tau \mid s, \boldsymbol{\theta}) = \sum_d g(\tau \mid \boldsymbol{\theta}_{d,s}) \prod_{d' \neq d} \{1 - G(\tau \mid \boldsymbol{\theta}_{d',s})\}. \tag{3.3}$$

The marginal probability of taking decision d under the influence of stimulus s is likewise obtained as

$$f(d \mid s, \boldsymbol{\theta}) = \int_{\delta_s}^{\infty} g(\tau \mid \boldsymbol{\theta}_{d,s}) \prod_{d' \neq d} \{1 - G(\tau \mid \boldsymbol{\theta}_{d',s})\}\, d\tau. \tag{3.4}$$

Interestingly, model (3.4) is similar to traditional multinomial probit/logit regression models (Borooah, 2002; Agresti, 2018) except that the latent variables are now inverse Gaussian distributed as opposed to being normal or extreme-value distributed, and the observed category is associated with the minimum of the latent variables in contrast to being identified with the maximum of the latent variables.

In an interesting recent work, Kunkel *et al.* (2019) have also used an inverse Gaussian distribution based hierarchical Bayesian model for decision making, albeit in a simpler binary category case, focusing primarily on individual level models with no mechanism to assess population level effects or their dynamic complexities.

For our motivating longitudinal tone learning experiment described in Section 3.2, for $i \in \{1, \ldots, n = 20\}, \ell \in \{1, \ldots, L = 40\}, t \in \{1, \ldots, T = 10\}$, let $s_{i,\ell,t}$ denote the input tone for the i^{th} individual in the ℓ^{th} trial in block t. Likewise, let $d_{i,\ell,t}$ and $\tau_{i,\ell,t}$ denote, respectively, the selected Mandarin tone and the time taken to reach the corresponding threshold by the i^{th} individual in the ℓ^{th} trial in block t. We now have

$$g\{\tau_{i,\ell,t} \mid s_{i,\ell,t} = s, \boldsymbol{\theta}_{d,s}^{(i)}(t)\} = \frac{b_{d,s}^{(i)}(t)}{\sqrt{2\pi}(\tau_{i,\ell,t} - \delta_s^{(i)})^{3/2}} \exp\left[-\frac{\{b_{d,s}^{(i)}(t) - \mu_{d,s}^{(i)}(t)(\tau_{i,\ell,t} - \delta_s^{(i)})\}^2}{2(\tau_{i,\ell,t} - \delta_s^{(i)})}\right]. \tag{3.5}$$

The drift rates $\mu_{d,s}^{(i)}(t)$ and the decision boundaries $b_{d,s}^{(i)}(t)$ now also vary with the blocks t. In addition, we accommodate random effects by allowing $\delta_s^{(i)}$, $\mu_{d,s}^{(i)}(t)$ and

$b_{d,s}^{(i)}(t)$ to also depend on the subject index i. We let $y_{i,\ell,t} = (d_{i,\ell,t}, \tau_{i,\ell,t})$, $\mathbf{y} = \{y_{i,\ell,t}\}_{i,\ell,t}$, and $d_0 = 4$ be the number of possible decision categories (T1, T2, T3, T4). The likelihood function of our longitudinal drift-diffusion mixed model thus takes the form

$$L(\mathbf{y} \mid \mathbf{s}, \boldsymbol{\theta}) = \prod_{d=1}^{d_0} \prod_{s=1}^{d_0} \prod_{t=1}^{T} \prod_{i=1}^{n} \prod_{\ell=1}^{L} \left(g\{\tau_{i,\ell,t} \mid \boldsymbol{\theta}_{d,s}^{(i)}(t)\} \prod_{d' \neq d} [1 - G\{\tau_{i,\ell,t} \mid \boldsymbol{\theta}_{d',s}^{(i)}(t)\}] \right)^{\mathbb{1}\{d_{i,\ell,t}=d, s_{i,\ell,t}=s\}}.$$

3.3.1 Modeling the Offsets

The offset parameters $\delta_s^{(i)}$, we recall, signify the times spent on encoding the different input tones, the time to press computer keys to record the responses, etc., and hence are not directly relevant to the actual decision making processes. These parameters are thus biologically not very interesting but may still vary between individuals and have an important effect on the estimates of drift rates and boundaries (Teichert *et al.*, 2016). We thus let them vary between input stimuli and participants but assume them to remain stable across blocks as in (3.5).

We assign uniform priors on $\delta_s^{(i)} \sim \text{Unif}(0, \delta_{s,i,\max})$, where $\delta_{s,i,\max}$ is the minimum of all response times under stimulus s for individual i, that is, $\delta_{s,i,\max} = \min_{\{(\ell,t):s_{i,\ell,t}=s\}} \tau_{i,\ell,t}$.

3.3.2 Modeling the Drifts and the Boundaries

Our modeling efforts concentrate henceforth on flexibly characterizing the longitudinal evolution of the mixed effects parameters $\mu_{d,s}^{(i)}(t), b_{d,s}^{(i)}(t)$. Variations in

these parameters over training blocks explain perceptual learning in the participants. Variations across participants, on the other hand, explain their performance heterogeneity. Following the discussion in the introduction, of particular interest are the local similarities and differences between these parameters for different input-response tone combinations (d, s) in different learning phases.

To this end, we propose essentially identical modeling strategies for $\mu_{d,s}^{(i)}(t)$ and $b_{d,s}^{(i)}(t)$. For ease of exposition avoiding unnecessary repetition, we describe below only these common strategies using simplified generic notations. With $x = (d, s) \in \mathcal{X} = \{(1,1), (1,2), \ldots, (4,4)\} \equiv \{1, 2, \ldots, x_{\max}\}$, $x_{\max} = 4 \times 4$, succinctly representing the input-response tone combinations and, with some abuse, $\theta_x^{(i)}(t)$ being a generic for $\mu_{d,s}^{(i)}(t)$ and $b_{d,s}^{(i)}(t)$, we let

$$\theta_x^{(i)}(t) = \exp\{f_x(t) + u_x^{(i)}(t)\}, \qquad u_x^{(i)}(t) \sim f_u\{u_x^{(i)}(t)\}. \tag{3.6}$$

The exponentiation in (3.6) enforces positivity constraints; $f_x(t)$ and $u_x^{(i)}(t)$ denote, respectively, additive fixed and random effects components in the exponential scale; f_u denotes the underlying random effects distribution. When needed, the fixed and random effects components for the drifts and the boundaries, as well as associated parameters and hyper-parameters, will be distinguished by reintroducing the subscripts as $f_{\mu,x}(t)$, $f_{b,x}(t)$, $u_{\mu,x}^{(i)}(t)$, $u_{b,x}^{(i)}(t)$ etc. To further simplify notation, generic data recording experimental blocks in $\{1, \ldots, T\}$ as well as other generic time points in $[1, T]$ will both be denoted by t. Likewise, generic input-response tone combinations as well as their particular values will both be denoted by x and so forth.

We model the components $f_x(t)$ and $u_x^{(i)}(t)$, and hence $\theta_x^{(i)}(t)$, to all be

smoothly varying functions over $t \in [1, T]$. A functional approach is not strictly necessary if inference is restricted only to the T data recording blocks blocks $t \in \{1, \ldots, T\}$. Learning may, however, be viewed as a continuous process - the brain synthesizes information from relevant past experiences even when not being actively engaged in actual decision making. A functional approach to modeling $f_x(t)$ and $u_x^{(i)}(t)$ for any $t \in [1, T]$, not just the experimental blocks $t \in \{1, \ldots, T\}$, thus facilitates parameter interpretability. A functional approach is also practically convenient in characterizing smoothly varying longitudinal parameter trajectories.

In modeling the fixed effects components $f_x(t)$, we are not only interested in characterizing their overall trajectories over time t for different input-response combinations $x = (d, s)$ but also how they might vary locally between different values of x in different learning stages. Compared to the fixed effects, we have to, however, rely on much less data to estimate the random effects $u_x^{(i)}(t)$ for different $x = (d, s)$ and different participant i, especially for $d \neq s$ toward later stages of the experiment when most participants identify the input tones with high accuracies. Our models and inferential goals for the random effects $u_x^{(i)}(t)$ will therefore be relatively modest.

3.3.2.1 Locally Varying Functional Fixed Effects

We now propose a novel approach to modeling the latent functions $f_x(t)$ using basis decomposition methods that allow them to smoothly vary with the blocks t while also depending locally on the indexing variable x. To begin with, we let

$$f_x(t) = \sum_{k=1}^{K} \beta_k^{(x)} B_k(t), \tag{3.7}$$

where $\mathbf{B}(t) = \{B_1(t), \ldots, B_K(t)\}^{\mathrm{T}}$ are a set of known locally supported basis functions spanning $[1, T]$, $\boldsymbol{\beta}^{(x)} = (\beta_1^{(x)}, \ldots, \beta_K^{(x)})^{\mathrm{T}}$ are associated unknown coefficients to be estimated from the data. In this work, we use quadratic B-spline bases with knot points coinciding with the block locations. B-splines are non-negative, continuous and have desirable local supports (Figure 2.2). Mixtures of B-splines are highly flexible (de Boor, 1978). Allowing the $\beta_k^{(x)}$'s to flexibly vary with x, the model can accommodate widely different shapes for different input-response tone combinations.

It is difficult to assess how similar or different these functions are using such unstructured models. One potential solution is to cluster the spline coefficients $\boldsymbol{\beta}^{(x)}$ associated with different input-response tone combinations x. If, for example, $\boldsymbol{\beta}^{(x_1)} = \boldsymbol{\beta}^{(x_2)}$ for two combinations x_1 and x_2, then we have $f_{x_1}(t) = f_{x_2}(t)$ for all t.

Such global clustering of all elements of $\boldsymbol{\beta}^{(x)}$ together does not, however, allow us to straightforwardly assess the local similarities and differences between these functions in different learning phases. To induce a desirable local cluster inducing mechanism, we introduce a set of latent variables $z_k^{(x)}$ for each input-response tone combination x with a shared state space $\mathcal{X}$, and associated core coefficients $\beta_{k,z}^{\star}$ and let

$$(\beta_k^{(x)} \mid z_k^{(x)} = z_k) = \beta_{k,z_k}^{\star},$$

implying

$$\{f_x(t) \mid z_k^{(x)} = z_k, k = 1, \ldots, K\} = \sum_{k=1}^{K} \beta_{k,z_k}^{\star} B_k(t). \tag{3.8}$$

The set of B-spline coefficients to be estimated at the k^{th} location now comprises the $\beta_{k,z_k}^{\star}$'s that are indexed by $z_k^{(x)} = z_k$ at that location k. When $z_k^{(x_1)} = z_k^{(x_2)}$

for two different levels x_1 and x_2 of x, we have $\beta_k^{(x_1)} = \beta_k^{(x_2)}$ and the implied functions $f_{x_1}(t)$ and $f_{x_2}(t)$ will tend to be similar at location k. Indeed, for quadratic B-splines with knots at the blocks $\{1, \ldots, T\}$, $f_{x_1}(t)$ and $f_{x_2}(t)$ will be exactly equal at block t when $z_t^{(x_1)} = z_t^{(x_2)}$ and $z_{t+1}^{(x_1)} = z_{t+1}^{(x_2)}$.

In theory, we could use B-splines of other small degrees as they all enjoy local support properties. With linear splines, however, smoothness becomes harder to control, and with cubic splines, three latent variables would be needed to determine the cluster configuration at each block t. We found quadratic B-splines to be a good compromise between the two for modeling smoothly varying curves while also maintaining easy interpretability of the latent variables.

Letting $\mathcal{Z}_k = \{z_k : z_k^{(x)} = z_k \text{ for some } x \in \mathcal{X}\}$, the case $|\mathcal{Z}_k| = 1$ then characterizes the scenario when the the spline coefficients for all input-response tone combinations x are the same at location k. On the other end, when $|\mathcal{Z}_k| = x_{\max} = 4 \times 4$, the spline coefficients are all different for different x at location k. In our tone learning application, $|\mathcal{Z}_k|$ tend to be much smaller than $x_{\max}$ uniformly for all k and the restricted support $z_k^{(x)} \in \{1, \ldots, z_{\max}\} \subset \mathcal{X}$ with $z_{\max} = 8 < x_{max} = 16$ will suffice.

We model the temporal evolution of the latent local cluster indicators $z_k^{(x)}, k = 1, \ldots, K$, using hidden Markov models (HMMs) (Figure 3.3). We consider two types of dynamics for the latent states corresponding to correct (C) and incorrect (I) iden-

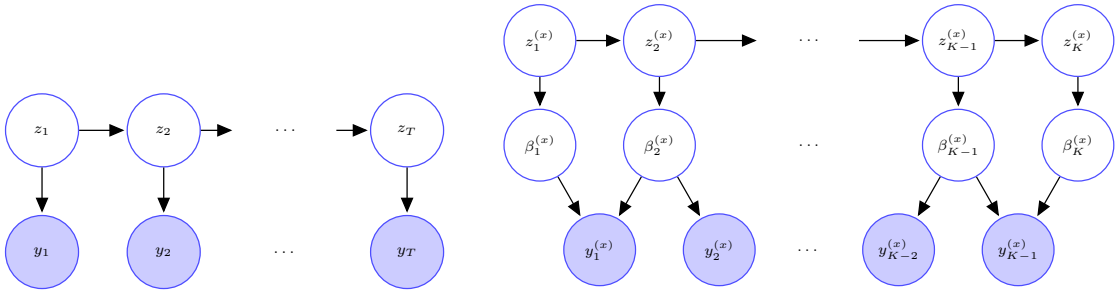


Figure 3.3: Left panel: Graph of a conventional HMM. Right panel: Graph of our proposed functional HMM model (3.8) with quadratic B-splines (Figure 2.2) with knots points coinciding with the data recording time blocks ($T = K - 1$).

tification of the tones. That is,

$$(z_k^{(d,s)} \mid z_{k-1}^{(d,s)} = z_{k-1}) \sim \text{Mult}(\pi^{(C)}_{z_{k-1},1}, \ldots, \pi^{(C)}_{z_{k-1},z_{\max}}) \quad \text{when } d = s,$$
$$(z_k^{(d,s)} \mid z_{k-1}^{(d,s)} = z_{k-1}) \sim \text{Mult}(\pi^{(I)}_{z_{k-1},1}, \ldots, \pi^{(I)}_{z_{k-1},z_{\max}}) \quad \text{when } d \neq s.$$

The latent cluster inducing variables $z_k^{(x)}$'s are shared between $f_{\mu,x}(t)$ and $f_{b,x}(t)$, reducing computational complexities while also facilitating model interpretability. We assign Dirichlet priors on the transition probabilities

$$\boldsymbol{\pi}_z^{(C)} = (\pi^{(C)}_{z,1}, \ldots, \pi^{(C)}_{z,z_{\max}})^{\mathrm{T}} \sim \text{Dir}(\alpha^{(C)}/z_{\max}, \ldots, \alpha^{(C)}/z_{\max}) \quad \text{with} \quad \alpha^{(C)} \sim \text{Ga}(a_\alpha, b_\alpha),$$
$$\boldsymbol{\pi}_z^{(I)} = (\pi^{(I)}_{z,1}, \ldots, \pi^{(I)}_{z,z_{\max}})^{\mathrm{T}} \sim \text{Dir}(\alpha^{(I)}/z_{\max}, \ldots, \alpha^{(I)}/z_{\max}) \quad \text{with} \quad \alpha^{(I)} \sim \text{Ga}(a_\alpha, b_\alpha).$$

We next consider priors for the atoms $\beta^\star_{k,z_k}$. Conditional on the $z_k^{(x)}$'s and the coefficients at the previous locations, for $k = 2, \ldots, K$, we construct the priors

sequentially as

$$\beta^{\star}_{k,z_k} \sim \begin{cases} \prod\limits_{\{z^{(x)}_{k-1}:\ x \in \mathfrak{X}^{(z_k)}_k\}} \text{Normal}\left(\beta^{\star}_{k-1,z^{(x)}_{k-1}}, \sigma^2_{\beta,1}\right) & \text{if } |\mathfrak{X}^{(z_k)}_k| > 0, \\ \text{Normal}(\mu_{\beta,0}, \sigma^2_{\beta,0}) & \text{otherwise,} \end{cases} \tag{3.9}$$

where $\mathfrak{X}^{(z_k)}_k = \{x : z^{(x)}_k = z_k\}$ is the set of values of x that, at the location k, are assigned the label z_k. In constructing the prior in this manner, we center the core coefficients around the ones that are 'expressed' at the previous location (Figure 3.4), penalizing their first order differences. The coefficients that are not associated with any levels of x are assigned a normal prior with a large variance $\sigma^2_{\beta,0}$. The initial coefficients are assigned non-informative flat priors as $\beta^{\star}_{1,z_k} \sim 1$.

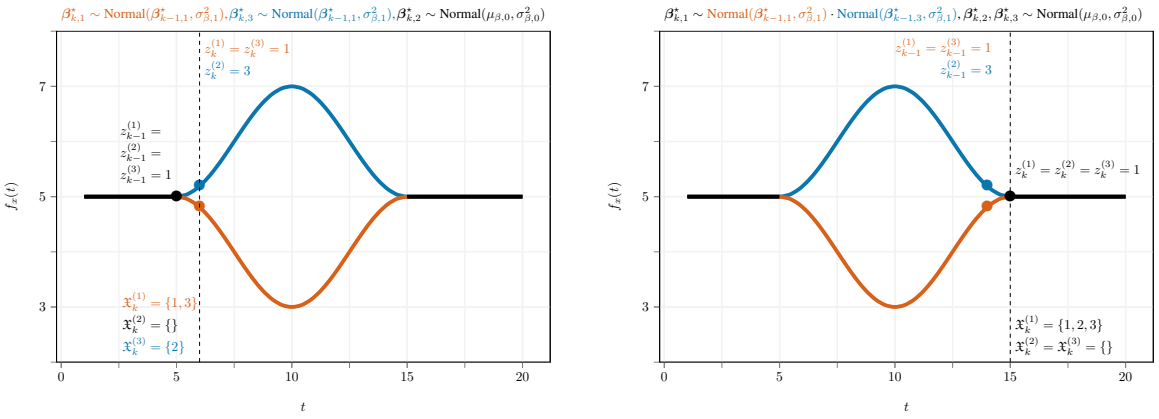


Figure 3.4: An illustration of the prior on the spline core coefficients $\beta^{\star}_{k,z_k}$ at location k (marked by the dashed vertical lines) in the fixed effects model developed in Section 3.3.2.1 for a synthetic scenario with $x \in \{1,2,3\}$, where the curves corresponding to the three levels of x are initially equal, the curves for $x = 1,3$ (in red) and $x = 2$ (in blue) then diverge at $t = 6$, merging back again at $t = 15$.

As an illustration on the smoothness inducing prior, consider the example in Figure 3.4. In the left panel, at location $k-1 = 5$, all of the levels for the covariate

x are assigned to the first latent state, yielding the same curve for the three levels of x. At location $k = 6$, levels 1 and 3 are assigned to the first latent state, whereas level 2 is assigned to the third latent state. This corresponds to the case in which the curves for $x = 1, 3$ and $x = 2$ diverge. Therefore, using (3.9),

- $\mathfrak{X}_k^{(1)} = \{x : z_k^{(x)} = 1\} = \{1, 3\}$ and the conditional prior for the core coefficient of the first latent state is $\beta_{k,1}^\star \sim \prod_{j \in \{z_{k-1}^{(1)}, z_{k-1}^{(3)}\}} \text{Normal}(\beta_{k-1,j}^\star, \sigma_{\beta,1}^2) = \text{Normal}(\beta_{k-1,1}^\star, \sigma_{\beta,1}^2)$,
- $\mathfrak{X}_k^{(2)} = \{x : z_k^{(x)} = 2\} = \emptyset$ and the conditional prior for the core coefficient of the second latent state is $\beta_{k,2}^\star \sim \text{Normal}(\mu_{\beta,0}, \sigma_{\beta,0}^2)$,
- $\mathfrak{X}_k^{(3)} = \{x : z_k^{(x)} = 3\} = \{2\}$ and the conditional prior for the core coefficient of the third latent state is $\beta_{k,3}^\star \sim \prod_{j \in \{z_{k-1}^{(2)}\}} \text{Normal}(\beta_{k-1,j}^\star, \sigma_{\beta,1}^2) = \text{Normal}(\beta_{k-1,1}^\star, \sigma_{\beta,1}^2)$.

In the right panel, at location $k - 1 = 14$, levels 1 and 3 are assigned to the first latent state, whereas level 2 is assigned to the third latent state. At location $k = 15$, all of the levels for the covariate x are assigned to the first latent state. This corresponds to the case in which the curves for $x = 1, 3$ and $x = 2$ merge back. Therefore,

- $\mathfrak{X}_k^{(1)} = \{x : z_k^{(x)} = 1\} = \{1, 2, 3\}$ and the conditional prior for the core coefficient of the first latent state is $\beta_{k,1}^\star \sim \prod_{j \in \{z_{k-1}^{(1)}, z_{k-1}^{(2)}, z_{k-1}^{(3)}\}} \text{Normal}(\beta_{k-1,j}^\star, \sigma_{\beta,1}^2) = \prod_{j \in \{1,3\}} \text{Normal}(\beta_{k-1,j}^\star, \sigma_{\beta,1}^2)$,

- $\mathfrak{X}_k^{(2)} = \{x : z_k^{(x)} = 2\} = \emptyset$ and the conditional prior for the core coefficient of the second latent state is $\beta_{k,2}^{\star} \sim \text{Normal}(\mu_{\beta,0}, \sigma_{\beta,0}^2)$,
- $\mathfrak{X}_k^{(3)} = \{x : z_k^{(x)} = 3\} = \emptyset$ and the conditional prior for the core coefficient of the third latent state is $\beta_{k,3}^{\star} \sim \text{Normal}(\mu_{\beta,0}, \sigma_{\beta,0}^2)$.

The smoothness of the curves is controlled by the parameter $\sigma_{\beta,1}^2$ and is assigned a prior, allowing it to be informed by the data. We let

$$\sigma_{\beta,1}^2 \sim \text{C}^+(0, 1),$$

where $\text{C}^+(a, b)$ denotes a half-Cauchy distribution (Gelman, 2006; Polson and Scott, 2012) with location parameter a and scale parameter b. The half-Cauchy distribution, which attains its mode at zero, is capable of capturing strong smoothness, while also having heavy tails, thus being capable of capturing wiggly functions. The choice of the scale hyper-parameter is discussed in Section 3.4.2.

Importantly, although our basic building blocks for the fixed effects components comprise conventional HMMs, one for each input-response tone combination $x = (d, s)$, for any input tone s, all four latent variables $z_k^{(1,s)}, z_k^{(2,s)}, z_k^{(3,s)}, z_k^{(4,s)}$ simultaneously appear in equation (3.2). For each input tone, the graph for our tone learning model (Figure 3.5 and Figure 3.7) thus resembles a factorial HMM (Ghahramani and Jordan, 1997, fHMM) with four hidden layers. In the posterior, a latent state $z_k^{(d,s)}$ is thus informed by all responses generated under the tone s, not just the subset corresponding to $x = (d, s)$. This has important consequences for posterior inference, as we discuss in Section 3.4.

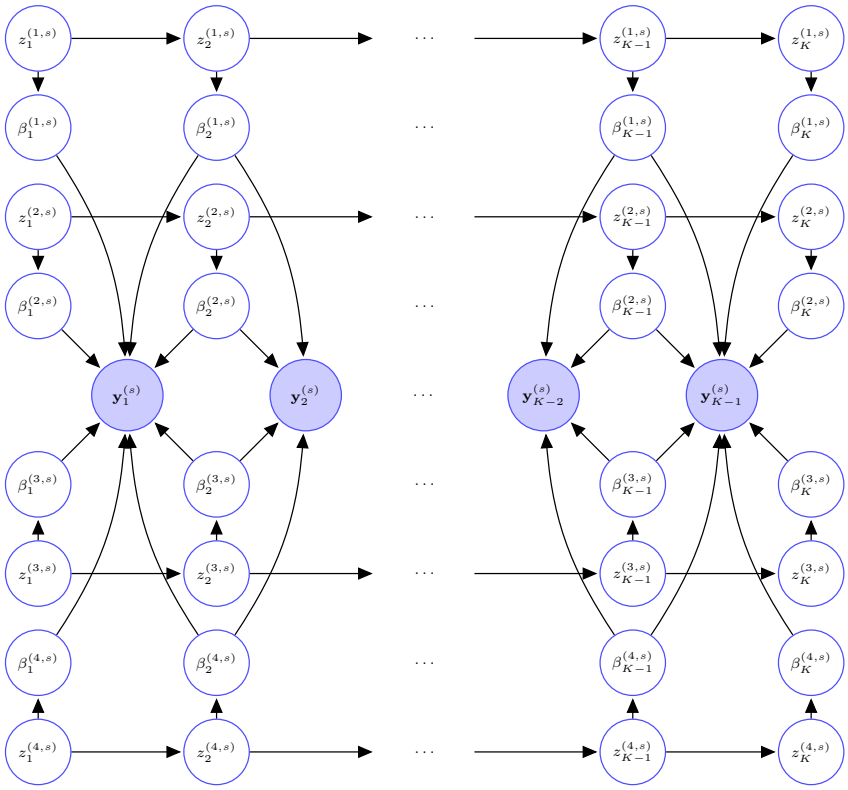


Figure 3.5: Graph of the proposed fixed effects model for tone learning.

3.3.2.2 Locally Varying Functional Random Effects

We now focus on flexibly modeling the functional random effects components. For reasons outlined before Section 3.3.2.1, estimating $u_x^{(i)}(t)$ for each different x is a challenging task. For any participant, the random effects for correct and incorrect identification of the tones may, however, be expected to be on the opposite sides of the corresponding population level curves. Taking a middle path, we thus allow different random effects $u_C^{(i)}(t)$ and $u_I^{(i)}(t)$ for correct (C) and incorrect (I) identifications,

respectively, as

$$u_{d,s}^{(i)}(t) = u_C^{(i)}(t) \quad \text{when } d = s, \qquad u_{d,s}^{(i)}(t) = u_I^{(i)}(t) \quad \text{when } d \neq s.$$

We adopt a common strategy to model both $u_C^{(i)}(t)$ and $u_I^{(i)}(t)$. Suppressing the subscripts to simplify notation and avoid repetition, we model the time-varying random effects components $u^{(i)}(t)$ as

$$\begin{aligned} u^{(i)}(t) &= \textstyle\sum_{k=1}^{K} \beta_{k,u}^{(i)} B_k(t), \\ \boldsymbol{\beta}_u^{(i)} &\sim \mathrm{MVN}_K\{\mathbf{0}, (\sigma_{u,a}^{-2}\mathbf{I}_K + \sigma_{u,s}^{-2}\mathbf{P}_u)^{-1}\}, \end{aligned} \tag{3.10}$$

where $\boldsymbol{\beta}_u^{(i)} = (\beta_{1,u}^{(i)}, \ldots, \beta_{K,u}^{(i)})^{\mathrm{T}}$ are subject-specific spline coefficients, $\mathrm{MVN}_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a K dimensional multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. We choose $\mathbf{P}_u = \mathbf{D}_u^{\mathrm{T}}\mathbf{D}_u$, where the $(K-1) \times K$ matrix $\mathbf{D}_u$ is such that $\mathbf{D}_u\boldsymbol{\beta}_u^{(i)}$ computes the first order differences in $\boldsymbol{\beta}_u^{(i)}$. The model thus penalizes $\sum_{k=1}^{K}(\nabla\beta_{k,u}^{(i)})^2 = \boldsymbol{\beta}_u^{(i)\mathrm{T}}\mathbf{P}_u\boldsymbol{\beta}_u^{(i)}$, the sum of squares of first order differences in $\boldsymbol{\beta}_u^{(i)}$ (Eilers and Marx, 1996). The random effects variance parameter $\sigma_{u,s}^2$ models the smoothness of the random effects curves, smaller $\sigma_{u,s}^2$ inducing smoother $u^{(i)}(t)$'s. Additional variations from the constant zero curve are explained by $\sigma_{u,a}^2$ (Figure 3.6). The absence of random effects is signified by the limiting case $\sigma_{u,s}^2 = \sigma_{u,a}^2 = 0$. We assign half-Cauchy priors on the variance parameters as

$$\sigma_{u,s}^2 \sim \mathrm{C}^+(0,1), \qquad \sigma_{u,a}^2 \sim \mathrm{C}^+(0,1).$$

Modeled in the same space of quadratic B-splines, the fixed and the random effects curves thus share similar smoothness properties. Having different smoothness

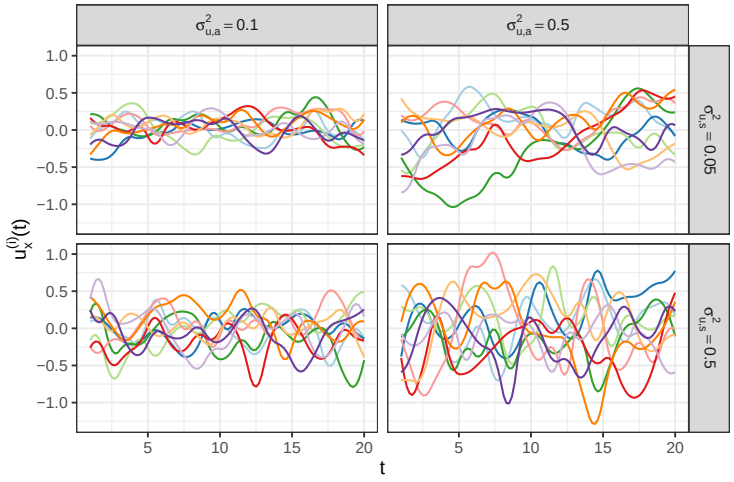


Figure 3.6: An illustration of the functional random effects model proposed in Section 3.3.2.2. Each panel shows a collection of 10 random draws from the random effects distribution for a combination of values of $(\sigma^2_{u,s}, \sigma^2_{u,a})$.

controlling parameters, they are, however, allowed to have different smoothness levels. A similar approach, but with additional assumptions on the covariance matrix of the random effects, has previously been developed in Guo (2002). To our knowledge, model (3.10) for the random effects is thus also novel to the literature.

Integrating out the random effects, the corresponding population level parameters $\theta_x(t)$ are obtained as

$$\theta_x(t) = \int \exp\{f_x(t) + u_x^{(i)}(t)\} f_u\{u_x^{(i)}(t)\} du_x^{(i)}(t) = \exp\left[f_x(t) + \frac{\text{var}\{u_x^{(i)}(t)\}}{2}\right].$$

3.4 Posterior Inference

Posterior inference for conventional HMMs can generally be based on samples drawn from the posterior using dynamic message passing MCMC algorithms (Rabiner, 1989; Scott, 2002). The nonstandard inverse Gaussian likelihood and the fHMM type model structure of our proposed longitudinal drift-diffusion mixed model, however, bring in significant additional complexities. We adapt recent advances in MCMC algorithms for discrete spaces (Neal, 2003; Van Gael *et al.*, 2008; Titsias and Yau, 2014; Zanella, 2019) in novel non-trivial ways, designing locally informative slice sampling moves that carefully exploit the conditional independence relationships encoded in the model to overcome the computational challenges.

3.4.1 MCMC Algorithm

Posterior inference for the longitudinal drift-diffusion mixed model, described in Section 3.3, is based on samples drawn from the posterior using a message passing MCMC algorithm.

In what follows, $\boldsymbol{\zeta}$ denotes a generic variable that collects all other variables not explicitly mentioned, including the data points. Also, p_0 will sometimes be used as a generic for a prior distribution without explicitly mentioning its hyperparameters. The sampler for the drift diffusion model of Section 3.3 comprises the following steps.

1. Update the offset parameters $\delta_s^{(i)}, s = 1, \ldots, d_0$. The full conditionals $p(\delta_s^{(i)} \mid \boldsymbol{\zeta}) \propto p_0(\delta_s^{(i)}) L(\mathbf{y} \mid \mathbf{s}, \boldsymbol{\theta})$ do not have closed forms. Metropolis-Hastings (MH) steps with

log-normal proposals centered on the previous sampled values are used to update these parameters.

2. Jointly update the drift and boundary spline coefficients $(\beta^{\star}_{\mu,k,z_k}, \beta^{\star}_{b,k,z_k})$, for $k = 1, \ldots, K$.

   (a) If the parameters are assigned to one of the clusters, the full conditionals do not have closed forms. MH steps are therefore used with the smoothness inducing priors (3.9) on $(\beta^{\star}_{\mu,k,z_k}, \beta^{\star}_{b,k,z_k})$ as the proposal distributions.

   (b) If the parameters are not assigned to any of the clusters, the full conditional distribution is the second term of the prior in (3.9).

3. Update the latent cluster assignments $\mathbf{z}_k^{(s)} = (z_k^{(1,s)}, \ldots, z_k^{(4,s)})^{\mathrm{T}}$:

   (a) Sample the auxiliary variables $\mathbf{v}_k^{(s)} = (v_k^{(1,s)}, \ldots, v_k^{(4,s)})^{\mathrm{T}}$ as

   $$\begin{aligned}
   &\text{for } k = 1, \ldots, K-1,\\
   &p(\mathbf{v}_k^{(s)} \mid \mathbf{z}_k^{(s)}, \mathbf{z}_{k+1}^{(s)}, \mathbf{y}_k^{(s)}, \boldsymbol{\zeta}) \propto g\{p(\mathbf{y}_k^{(s)} \mid \mathbf{v}_k^{(s)}, \mathbf{z}_{k+1}^{(s)}, \boldsymbol{\zeta})\} \mathbb{1}\{\mathbf{v}_k^{(s)} \in \mathcal{H}_m(\mathbf{z}_k^{(s)})\},\\
   &p(\mathbf{v}_K^{(s)} \mid \mathbf{z}_K^{(s)}, \boldsymbol{\zeta}) \propto \mathbb{1}\{\mathbf{v}_K^{(s)} \in \mathcal{H}_m(\mathbf{z}_K^{(s)})\}.
   \end{aligned}$$

   (b) Back-propagate the messages $m_k(\mathbf{z}_k^{(s)}) = p(\mathbf{y}_{k:(K-1)}^{(s)}, \mathbf{v}_{k:K}^{(s)} \mid \mathbf{z}_k^{(s)}, \boldsymbol{\zeta})$ using the

recursion

$$
\begin{aligned}
m_k(\mathbf{z}_k^{(s)}) &= p(\mathbf{y}_{k:(K-1)}^{(s)}, \mathbf{v}_{k:K}^{(s)} \mid \mathbf{z}_k^{(s)}, \boldsymbol{\zeta}) \\
&= \sum_{\mathbf{z}_{k+1}^{(s)}} p(\mathbf{y}_{k:(K-1)}^{(s)}, \mathbf{v}_{k:K}^{(s)} \mid \mathbf{z}_k^{(s)}, \mathbf{z}_{k+1}^{(s)}, \boldsymbol{\zeta}) p(\mathbf{z}_{k+1}^{(s)} \mid \mathbf{z}_k^{(s)}, \boldsymbol{\zeta}) \\
&= \sum_{\mathbf{z}_{k+1}^{(s)}} p(\mathbf{y}_k^{(s)}, \mathbf{v}_k^{(s)} \mid \mathbf{z}_k^{(s)}, \mathbf{z}_{k+1}^{(s)}, \boldsymbol{\zeta}) p(\mathbf{y}_{(k+1):(K-1)}^{(s)}, \mathbf{v}_{(k+1):K}^{(s)} \mid \mathbf{z}_k^{(s)}, \mathbf{z}_{k+1}^{(s)}, \boldsymbol{\zeta}) p(\mathbf{z}_{k+1}^{(s)} \mid \mathbf{z}_k^{(s)}, \boldsymbol{\zeta}) \\
&= \sum_{\mathbf{z}_{k+1}^{(s)}} p(\mathbf{y}_k^{(s)}, \mathbf{v}_k^{(s)} \mid \mathbf{z}_k^{(s)}, \mathbf{z}_{k+1}^{(s)}, \boldsymbol{\zeta}) p(\mathbf{y}_{(k+1):(K-1)}^{(s)}, \mathbf{v}_{(k+1):K}^{(s)} \mid \mathbf{z}_{k+1}^{(s)}, \boldsymbol{\zeta}) p(\mathbf{z}_{k+1}^{(s)} \mid \mathbf{z}_k^{(s)}, \boldsymbol{\zeta}) \\
&= \sum_{\mathbf{z}_{k+1}^{(s)}} p(\mathbf{v}_k^{(s)} \mid \mathbf{z}_k^{(s)}, \mathbf{z}_{k+1}^{(s)}, \mathbf{y}_k^{(s)}, \boldsymbol{\zeta}) p(\mathbf{y}_k^{(s)} \mid \mathbf{z}_k^{(s)}, \mathbf{z}_{k+1}^{(s)}, \boldsymbol{\zeta}) p(\mathbf{z}_{k+1}^{(s)} \mid \mathbf{z}_k^{(s)}, \boldsymbol{\zeta}) m_{k+1}(\mathbf{z}_{k+1}^{(s)}), \\
&\propto \sum_{\mathbf{z}_{k+1}^{(s)} \in \mathcal{H}_m(\mathbf{v}_{k+1}^{(s)})} g\{p(\mathbf{y}_k^{(s)} \mid \mathbf{v}_k^{(s)}, \mathbf{z}_{k+1}^{(s)}, \boldsymbol{\zeta})\} \mathbb{1}\{\mathbf{v}_k^{(s)} \in \mathcal{H}_m(\mathbf{z}_k^{(s)})\} p(\mathbf{y}_k^{(s)} \mid \mathbf{z}_k^{(s)}, \mathbf{z}_{k+1}^{(s)}, \boldsymbol{\zeta}) \\
&\qquad\qquad p(\mathbf{z}_{k+1}^{(s)} \mid \mathbf{z}_k^{(s)}, \boldsymbol{\zeta}) m_{k+1}(\mathbf{z}_{k+1}^{(s)}),
\end{aligned}
$$

starting with the final condition $m_K(\mathbf{z}_K^{(s)}) = \mathbb{1}\{\mathbf{z}_K^{(s)} \in \mathcal{H}_m(\mathbf{v}_K^{(s)})\}$.

(c) Sample the latent cluster assignments forward one step at a time from

$$
\begin{aligned}
p(\mathbf{z}_{1:K}^{(s)} \mid \mathbf{y}_{1:(K-1)}^{(s)}, \mathbf{v}_{1:K}^{(s)}, \boldsymbol{\zeta}) = p(\mathbf{z}_K^{(s)} \mid \mathbf{z}_{1:(K-1)}^{(s)}, \mathbf{y}_{1:(K-1)}^{(s)}, \mathbf{v}_{1:K}^{(s)}, \boldsymbol{\zeta}) \dots \\
p(\mathbf{z}_1^{(s)} \mid \mathbf{y}_{1:(K-1)}^{(s)}, \mathbf{v}_{1:K}^{(s)}, \boldsymbol{\zeta}),
\end{aligned}
$$

where

$$
\begin{aligned}
&p(\mathbf{z}_k^{(s)} \mid \mathbf{z}_{1:(k-1)}^{(s)}, \mathbf{y}_{1:(K-1)}^{(s)}, \mathbf{v}_{1:K}^{(s)}, \boldsymbol{\zeta}) \propto p(\mathbf{y}_{k:(K-1)}^{(s)}, \mathbf{v}_{k:K}^{(s)} \mid \mathbf{z}_{1:k}^{(s)}, \boldsymbol{\zeta}) p(\mathbf{z}_k^{(s)} \mid \mathbf{z}_{1:(k-1)}^{(s)}, \boldsymbol{\zeta}) \\
&= p(\mathbf{y}_{1:(k-2)}^{(s)}, \mathbf{v}_{1:(k-2)}^{(s)} \mid \mathbf{z}_{1:k}^{(s)}, \boldsymbol{\zeta}) p(\mathbf{y}_{k-1}^{(s)}, \mathbf{v}_{k-1}^{(s)} \mid \mathbf{z}_{k-1}^{(s)}, \mathbf{z}_k^{(s)}, \boldsymbol{\zeta}) p(\mathbf{y}_{k:(K-1)}^{(s)}, \mathbf{v}_{k:K}^{(s)} \mid \mathbf{z}_k^{(s)}, \boldsymbol{\zeta}) \\
&\qquad\qquad p(\mathbf{z}_k^{(s)} \mid \mathbf{z}_{1:(k-1)}^{(s)}, \boldsymbol{\zeta}) \\
&\propto p(\mathbf{y}_{k-1}^{(s)}, \mathbf{v}_{k-1}^{(s)} \mid \mathbf{z}_{k-1}^{(s)}, \mathbf{z}_k^{(s)}, \boldsymbol{\zeta}) p(\mathbf{z}_k^{(s)} \mid \mathbf{z}_{1:(k-1)}^{(s)}, \boldsymbol{\zeta}) m_k(\mathbf{z}_k^{(s)}) \\
&= p(\mathbf{v}_{k-1}^{(s)} \mid \mathbf{z}_{k-1}^{(s)}, \mathbf{z}_k^{(s)}, \mathbf{y}_{k-1}^{(s)}, \boldsymbol{\zeta}) p(\mathbf{y}_{k-1}^{(s)} \mid \mathbf{z}_{k-1}^{(s)}, \mathbf{z}_k^{(s)}, \boldsymbol{\zeta}) p(\mathbf{z}_k^{(s)} \mid \mathbf{z}_{k-1}^{(s)}, \boldsymbol{\zeta}) m_k(\mathbf{z}_k^{(s)}) \\
&\propto g\{p(\mathbf{y}_{k-1}^{(s)} \mid \mathbf{v}_{k-1}^{(s)}, \mathbf{z}_k^{(s)}, \boldsymbol{\zeta})\} p(\mathbf{y}_{k-1}^{(s)} \mid \mathbf{z}_{k-1}^{(s)}, \mathbf{z}_k^{(s)}, \boldsymbol{\zeta}) p(\mathbf{z}_k^{(s)} \mid \mathbf{z}_{k-1}^{(s)}, \boldsymbol{\zeta}) m_k(\mathbf{z}_k^{(s)}).
\end{aligned}
$$

4. Update the cluster specific fixed effects spline coefficients:

$$
\begin{aligned}
(\beta_{\mu,k}^{(x)} \mid z_k^{(x)} = z_k, \boldsymbol{\zeta}) &\sim \mathbb{1}\{\beta_{\mu,k}^{(x)} = \beta_{\mu,k,z_k}^{\star}\}, \quad k = 1, \ldots, K. \\
(\beta_{b,k}^{(x)} \mid z_k^{(x)} = z_k, \boldsymbol{\zeta}) &\sim \mathbb{1}\{\beta_{b,k}^{(x)} = \beta_{b,k,z_k}^{\star}\}, \quad k = 1, \ldots, K.
\end{aligned}
$$

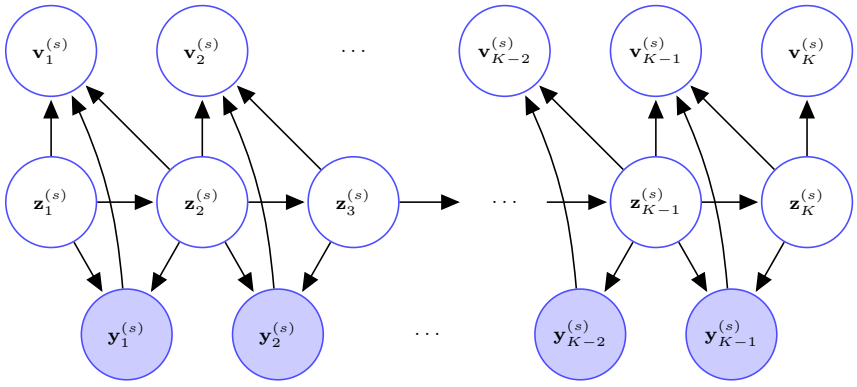


Figure 3.7: Locally informed Hamming ball sampling of the latent states in our tone-learning longitudinal drift-diffusion mixed model. See also Figure 3.5.

5. Update the transition probability matrices:

$$\left(\boldsymbol{\pi}_z^{(C)} \mid \boldsymbol{\zeta}\right) \sim \text{Dir}(\alpha^{(C)}/z_{\max} + n_{z,1}^{(C)}, \ldots, \alpha^{(C)}/z_{\max} + n_{z,z_{\max}}^{(C)})$$
$$\left(\boldsymbol{\pi}_z^{(I)} \mid \boldsymbol{\zeta}\right) \sim \text{Dir}(\alpha^{(I)}/z_{\max} + n_{z,1}^{(I)}, \ldots, \alpha^{(I)}/z_{\max} + n_{z,z_{\max}}^{(I)}),$$

where $n_{z,z'}^{(C)} = \sum_k \mathbb{1}\{z_k^{(x)} = z, z_{k+1}^{(x)} = z'\}$ is the number of transitions from z to z' for the HMMs associated with the correct identification of the tones, that is, with x s.t. $d = s$. A similar definition holds for $n_{z,z'}^{(I)}$.

6. Update the cluster specific smoothness parameter

$$p(\sigma^2_{\beta_\mu,1} \mid \boldsymbol{\zeta}) \propto \left(\sigma_{\beta_\mu,1}\right)^{-Kx_{\max}} \exp\left(-\frac{1}{2\sigma^2_{\beta_\mu,1}} \sum_x \boldsymbol{\beta}_\mu^{(x)\mathrm{T}} \mathbf{P}_u \boldsymbol{\beta}_\mu^{(x)}\right) p_0(\sigma^2_{\mu,u,a}).$$

MH steps with log-normal proposals centered on the previous sampled values are used to update these parameters.

7. Update the random effects spline coefficients $\beta_{\mu,k,u}^{(i)}$ and $\beta_{b,k,u}^{(i)}$: The full conditional does not have a closed form. An MH step with a normal proposal centered on the previous value was used.

8. Update the random effects variance parameters $\sigma^2_{\mu,u,a}$, $\sigma^2_{\mu,u,s}$, $\sigma^2_{b,u,a}$ and $\sigma^2_{b,u,s}$:
The full conditional for $\sigma^2_{\mu,u,a}$ is given by

$$p(\sigma^2_{\mu,u,a} \mid \boldsymbol{\zeta}) \propto \det(\sigma^{-2}_{\mu,u,s}\mathbf{P}_u + \sigma^{-2}_{\mu,u,a}\mathbf{I}_K)^{n/2} \exp\left(-\frac{1}{2\sigma^2_{\mu,u,a}} \sum_{i=1}^{n} \boldsymbol{\beta}_{\mu,u}^{(i)\mathrm{T}} \mathbf{P}_u \boldsymbol{\beta}_{\mu,u}^{(i)}\right) p_0(\sigma^2_{\mu,u,a}).$$

Analogous expressions can be found for the full conditionals of $\sigma^2_{\mu,u,s}$, $\sigma^2_{b,u,a}$ and $\sigma^2_{b,u,s}$. MH steps with log-normal proposals centered on the previous sampled values are used to update these parameters.

The main challenge here arises from the nonconjugacy of the inverse Gaussian distribution based likelihood function, requiring MH steps for updating $\delta_s^{(i)}, \beta^{\star}_{b,k,z_k}$ and $\beta^{\star}_{\mu,k,z_k}$. We employed the adaptive MH algorithm (Roberts and Rosenthal, 2009) for updating $\delta_s^{(i)}$ and the variance parameters, avoiding the difficult task of choosing the parameters of their proposal distributions while also improving mixing. Specifically, for every batch of 50 iterations, we inflate or deflate the standard deviation of the proposal distribution such that the optimal acceptance rate of 44% is achieved (Roberts *et al.*, 2001). The adaptive MH could not be employed for the cluster specific parameters $(\beta^{\star}_{b,k,z_k}, \beta^{\star}_{\mu,k,z_k})$ due to label switching, so we used tempered MH steps instead. For the proposal distributions for $(\beta^{\star}_{b,k,z_k}, \beta^{\star}_{\mu,k,z_k})$, we used the smoothness inducing conditional prior distributions $p_0(\beta^{\star}_{\mu,k,z_k} \mid \boldsymbol{\beta}^{\star}_{\mu,k-1}) \times p_0(\beta^{\star}_{b,k,z_k} \mid \boldsymbol{\beta}^{\star}_{b,k-1})$. Since the conditioning variables $\boldsymbol{\beta}^{\star}_{\mu,k-1}$ and $\boldsymbol{\beta}^{\star}_{b,k-1}$ are also updated at every iteration, the values sampled from the smoothness inducing priors are frequently accepted.

Based on M thinned samples $\{\boldsymbol{\theta}^{(m)}\}_{m=1}^{M}$ drawn from the posterior after the burn-in, the individual level drift parameters in the drift-diffusion mixed model are estimated as

$$\mu_x^{(i)}(t) = \exp\{f_{\mu,x}(t) + u_{\mu}^{(i)}(t)\} = \tfrac{1}{M}\sum_{m=1}^{M} \exp\{\widehat{f}_{\mu,x}^{(m)}(t) + \widehat{u}_{\mu}^{(i,m)}(t)\},$$

where $\widehat{f}_{\mu,x}^{(m)}(t) = \sum_{k=1}^{K} \beta^{\star(m)}_{\mu,k,z_k^{(x,m)}} B_k(t)$, $\widehat{u}_{\mu}^{(i,m)}(t) = \sum_{k=1}^{K} \beta_{k,u,\mu}^{(i,m)} B_k(t)$ etc. The population level drift parameters are likewise estimated as

$$\mu_x(t) = \int \exp\{f_{\mu,x}(t) + u_{\mu}^{(i)}(t)\} f\{u_{\mu}^{(i)}(t)\} du_{\mu}^{(i)}(t) = \exp\left[f_{\mu,x}(t) + \frac{\mathrm{var}\{u_{\mu}^{(i)}(t)\}}{2}\right]$$
$$= \frac{1}{M}\sum_{m=1}^{M} \exp\left\{\widehat{f}_{\mu,x}^{(m)}(t) + \frac{\mathrm{var}\{\widehat{u}_{\mu}^{(i,m)}(t)\}}{2}\right\}.$$

3.4.2 Prior Hyper-parameters and MCMC Initializations

The fixed effects parameters of the drift-diffusion mixed effects model (3.6) are initialized with an empirical Bayes type approach. As discussed in Section 3.3, the boundary and the drift parameters are related to the first two moments of the response times. Thus, we can use the empirical distribution of the response times to choose the initial guess for both drift and boundary parameters for each combination of input stimulus and response. The random effects are instead initialized at zero. The clustering configuration is initialized with all the success curves in different clusters, and all the failure curves in the same cluster.

Other crucial hyper-parameters are the mean and the standard deviation for the prior term of the unassigned components of $\boldsymbol{\beta}_{\mu}^{(x)}$ and $\boldsymbol{\beta}_{b}^{(x)}$, that is, the second term in the prior (3.9). We use the empirical distributions of the response times at every time point to set $\mu_{\beta,0}, \sigma^2_{\beta,0}$. The hyper-parameters in the Gamma(a_α, b_α) prior for the concentration parameters $\alpha^{(C)}$ and $\alpha^{(I)}$ of the Dirichlet distributions characterizing the latent variable dynamics are set at $a_\alpha = b_\alpha = 1$, as recommended in Escobar and West (1995). The half-Cauchy priors $\mathrm{C}^+(0, 1)$ on the smoothness parameters are non-informative for the smoothness of the corresponding longitudinal curves. The $\mathrm{C}^+(0, 1)$ distribution attains its mode at zero and hence is capable of capturing strong smoothness but also has heavy tails and is thus also capable of capturing wiggly functions. The left panel of Figure 3.8 shows some draws from $\mu_x(t) \mid \sigma^2_{\beta_\mu,1}$ with independent draws of the corresponding smoothness controlling parameter $\sigma^2_{\beta_\mu,1}$ from a $\mathrm{C}^+(0, 1)$ prior. A wide variety of curves are clearly sampled - some very smooth, some very wiggly, and many in between. Also, as the right panel

of Figure 3.8 illustrates, the posterior distributions of the smoothness parameters in our model all concentrate well within a region of flat $C^+(0,1)$ prior probability density. This is additional evidence that our prior is not producing any consistent bias in the posterior estimates.

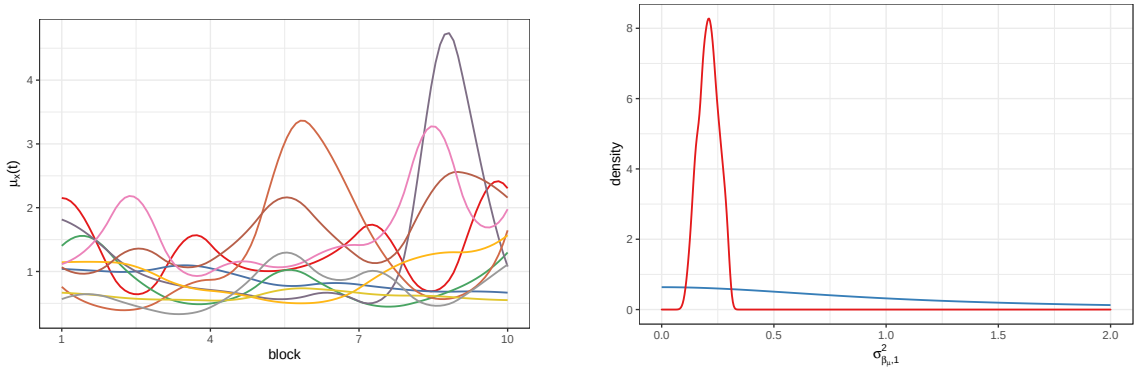


Figure 3.8: Left: 10 conditionally independent draws from $\mu_x(t) \mid \sigma^2_{\beta_\mu,1}$ with independent draws of $\sigma^2_{\beta_\mu,1}$ from a $C^+(0,1)$ prior. Right: The $C^+(0,1)$ prior distribution (in blue) and the corresponding posterior distribution (in red) for the smoothness parameter $\sigma^2_{\beta_\mu,1}$.

3.4.3 Convergence Diagnostics

This section presents some convergence diagnostics for the MCMC sampler. The results presented here are for the tone learning data set. Diagnostics for the simulation experiments were similar and hence omitted.

Figure 3.9 shows the trace plots of some individual level parameters at different training blocks. Figure 3.10 shows the trace plots of some individual level offset parameters. These results are based on the MCMC thinned samples. As these figures show, the running means are very stable and there seems to be no convergence issues.

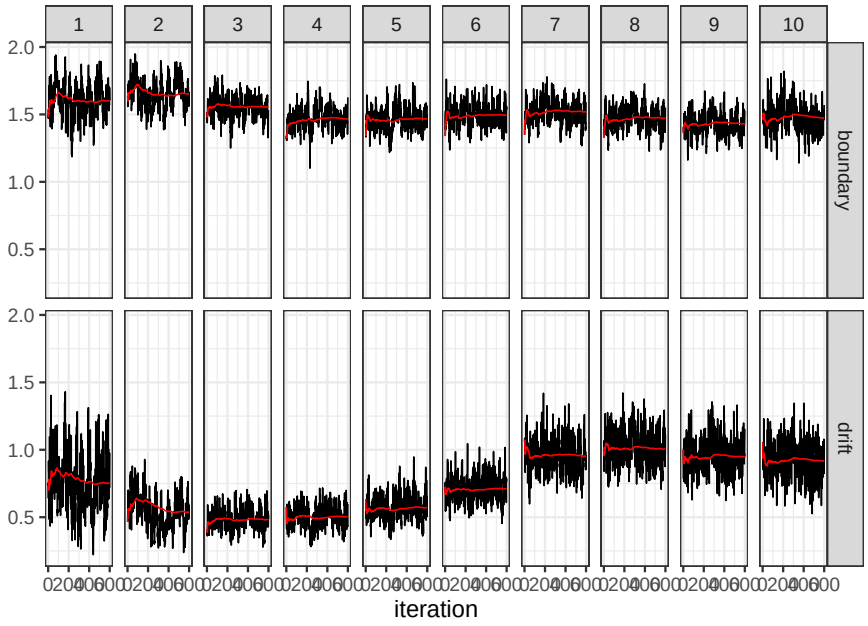


Figure 3.9: Trace plots of the individual drift rates $\mu_{1,1}^{(i)}(t)$ and boundary parameters $b_{1,1}^{(i)}(t)$ corresponding to the success categorization of tone T1 evaluated at each of the training blocks. The two rows correspond to the two different classes of parameters, and the ten columns to the training blocks. In each panel, the solid red line shows the running mean. Results for other drift and boundary parameters were very similar.

Additionally, the Geweke test (Geweke, 1991b) for stationarity of the chains, which formally compares the means of the first and last part of a Markov chain, was also performed. If the samples are drawn from the stationary distribution of the chain, the two means are equal and Geweke's statistic has an asymptotically standard normal distribution. The results of the test, reported in Table 3.1 and Table 3.2, indicate that convergence was satisfactory for the parameters considered. Only one parameter, $\mu_{1,1}^{(i)}(2)$ in the second row of Table 3.1, had a significant p-value. Some chance rejections are expected in multiple hypothesis testing scenarios. A visual inspection

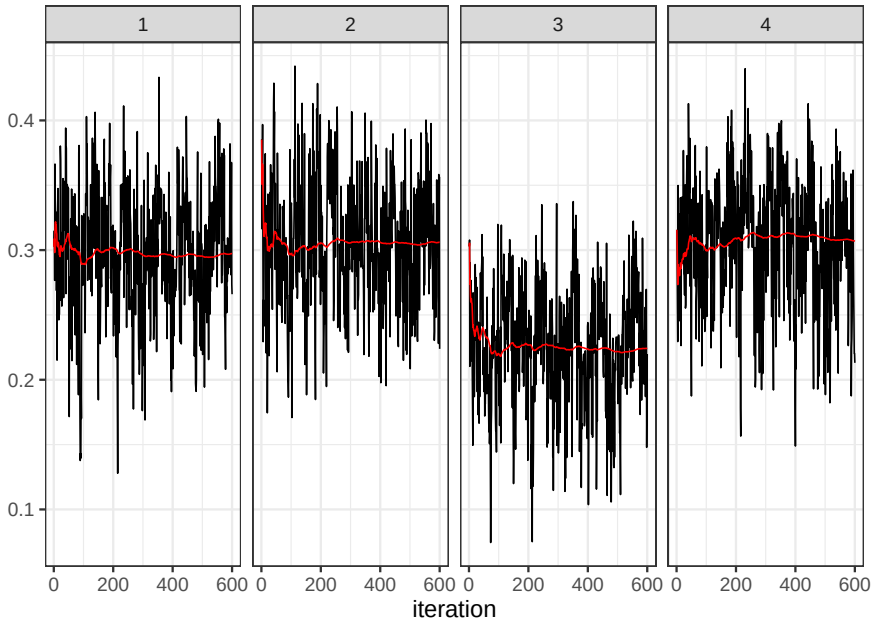


Figure 3.10: Trace plots of the individual level offset parameters $\delta_s^{(i)}$ for the four possible input tones. The four columns correspond to the input stimuli s. In each panel, the solid red line shows the running mean. Results for other offset parameters were very similar.

of the corresponding trace plot, however, does not indicate any serious issue.

	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	$t=7$	$t=8$	$t=9$	$t=10$
boundary	1.161	0.973	1.162	-1.287	-1.080	-0.554	0.164	-0.285	0.481	0.894
	(0.25)	(0.33)	(0.25)	(0.20)	(0.28)	(0.58)	(0.87)	(0.78)	(0.63)	(0.37)
drift	1.884	3.467	-0.102	-0.863	-1.171	-0.845	0.445	0.821	0.362	0.607
	(0.06)	(0.00)	(0.92)	(0.39)	(0.24)	(0.40)	(0.66)	(0.41)	(0.72)	(0.54)

Table 3.1: Geweke statistics and associated p-values assessing convergence of the individual drift rates $\mu_{1,1}^{(i)}(t)$ and boundary parameters $b_{1,1}^{(i)}(t)$ corresponding to the success categorization of tone T1 evaluated at each of the training blocks. Results for other drift and boundary parameters were very similar.

$s=1$	$s=2$	$s=3$	$s=4$
-0.395	-0.848	-0.019	-0.217
(0.69)	(0.40)	(0.98)	(0.83)

Table 3.2: Geweke statistics and associated p-values assessing convergence of the of the individual level offset parameters $\delta_s^{(i)}$ for the four possible input tones. Results for other offset parameters were very similar.

3.4.4 Software, Runtime, etc.

The results reported in this chapter are all based on $5,000$ MCMC iterations with the initial $2,000$ iterations discarded as burn-in. The remaining samples were further thinned by an interval of 5. We programmed in `R` and `C++`. The MCMC algorithm takes 10 hours on a Dell machine with 16 Gb RAM.

3.5 Application to Tone Categorization Data

In this section, we discuss the results produced by our method applied to the tone category learning data described in Section 3.2. Our primary inference goals, we recall, include understanding systematic longitudinal variations in perceptual categorization decision as the participants get better at identifying the four Mandarin tones with there being some additional interests in assessing individual specific trajectories, especially how they differ between good and bad learners.

Figure 3.11 shows the posterior mean trajectories and associated 90% credible intervals for the boundaries $b_{d,s}(t)$ and the drift rates $\mu_{d,s}(t)$ estimated by our method for different combinations of (d, s). Figure 3.12 reports the estimated posterior probabilities of each of the $\binom{4}{2} = 6$ pairs of success $(d = s)$ parameters to cluster

together in different blocks. These results suggest that after an initial learning phase, where the underlying processes are all similar across all input tones, there are two main learning groups. Two of the tones {T1, T3} seem to be easier to learn, as the corresponding drift parameters are larger, and tones {T2, T4} are more challenging. These findings are corroborated by empirical evidence and have significant biological relevance. The similarity groups of the mandarin tones are in fact {T1, T3}, which are characterized by the height of the pitch, and tones {T2, T4}, which are characterized by the direction of the pitch and are more challenging to learn. Tone T3, in particular, has a unique 'dipping' pitch pattern that is rarely encountered in English (Song *et al.*, 2008), and therefore is easier to categorize. Our proposed method allows similar inferential questions to be answered for the drift parameters corresponding to misclassifications, as well as for all the boundary parameters. The misclassification drift curves are mostly similar to each other, although some minor local differences can be found. Notable exceptions are $\mu_{1,3}(t)$ and $\mu_{3,1}(t)$ which are significantly smaller than all other drifts after the third block. As the participants get trained and experienced, for input tone T1, evidence in favor of tone T3 is thus collected more slowly compared to evidence in favor of T2 and T4, and vice versa. Likewise, while the boundary curve estimates mostly remain constant over the training blocks and similar to each other, $b_{1,3}(t)$ and $b_{3,1}(t)$ again differ from the rest and actually increase over the blocks. As the participants get trained and experienced, more evidence in favor of tone T3 is thus needed to misclassify tone T1 as tone T3 and vice versa. These suggest that, as the participants get trained and experienced, tones T1 and T3 become harder to misclassify for one another.

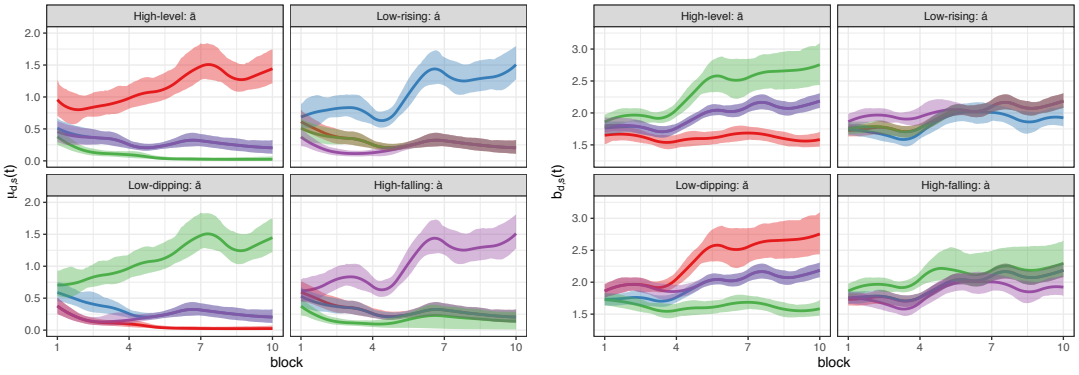


Figure 3.11: Results for tone learning data: Estimated posterior mean trajectories of the population level drifts $\mu_{d,s}(t)$ (left panel) and boundaries $b_{d,s}(t)$ (right panel) for the proposed longitudinal inverse Gaussian drift-diffusion mixed model. The shaded areas represent the corresponding 90% point wise credible intervals. Parameters for the high-level tone response category T1 are shown in red; low-rising T2 in blue; low-dipping T3 in green; and high-falling T4 in purple.

Importantly, our proposed drift-diffusion mixed model not only allows population level inference about the underlying processes but also allows us to assess individual specific parameter trajectories. Figure 3.13 shows the posterior mean trajectories and associated 90% credible intervals for the drift rates $\mu_{s,d}^{(i)}$ and the boundaries $b_{s,d}^{(i)}$ estimated by our method for the different success combinations of (d, s) for two participants - the one with the best accuracy averaged across all blocks, and the one with the worst accuracy averaged across all blocks. These results suggest significant individual specific heterogeneity. Importantly, the differences in the performances can again be explained mostly by differences in the drift trajectories. For the well performing participant, the drift trajectories increase rapidly with the training blocks before plateauing down around block 6 at which stage the participant has already attained native-like proficiency. For the poorly performing candidate on the

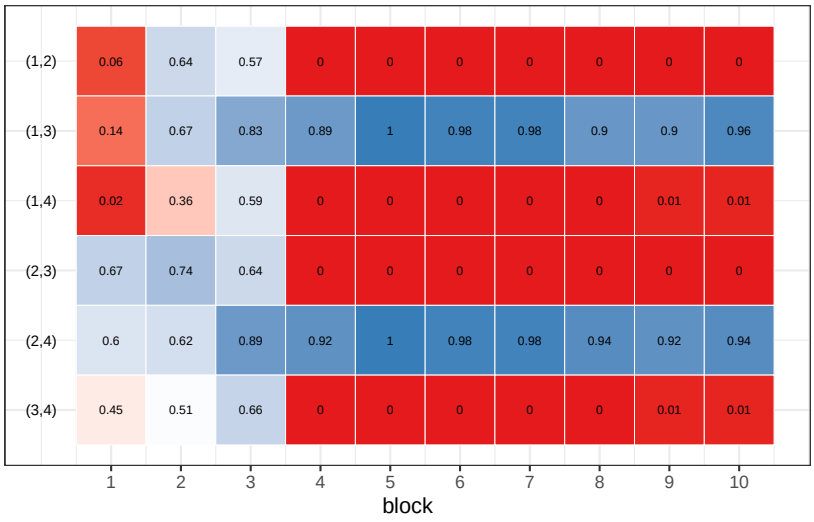


Figure 3.12: Results for tone learning data: Pairwise posterior co-clustering probabilities of the parameter trajectories for successful identification ($d = s$) of different input tones in different learning phases. The estimated posterior probability of $(\mu_{2,2}, b_{2,2})$ and $(\mu_{3,3}, b_{3,3})$ being clustered together, and hence being equal, in the 3^{th} block is thus 0.64, as shown in row $(2, 3)$ and column 3. Equivalently, the estimated posterior probability of $(\mu_{2,2}, b_{2,2})$ and $(\mu_{3,3}, b_{3,3})$ being different in the 3^{th} block is 0.36.

other hand, the drift trajectories remain approximately constant across all 10 blocks.

We compare the performance of our method with that of the linear ballistic accumulator (LBA) model (Brown and Heathcote, 2008), discussed in detail in Section 3.6. Similar to our model, the LBA uses independent evidence accumulators starting at δ that continue until a response threshold b is reached. The accumulator that first reaches the boundary corresponds to the decision outcome, and the time taken to reach this decision boundary is the observed response time. The LBA model, however, assumes that the evidence accumulates linearly at the rate μ, reaching the boundary b precisely at time $\tau = b/\mu$. Unlike in drift-diffusion models, where trial-

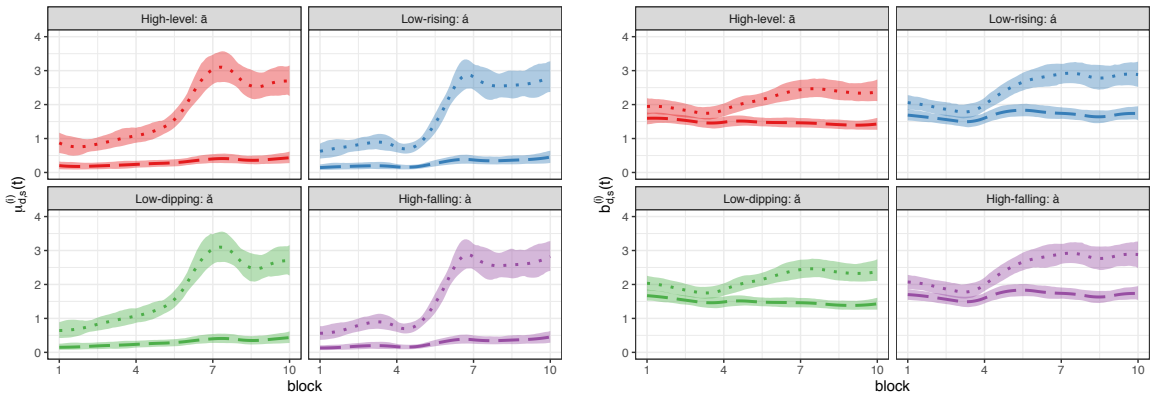


Figure 3.13: Results for tone learning data: Estimated posterior mean trajectories for individual specific drifts $\mu_{d,s}^{(i)}(t) = \exp\{f_{\mu,d,s}(t) + u_{\mu,C}^{(i)}(t)\}$ (left panel) and boundaries $b_{d,s}^{(i)}(t) = \exp\{f_{b,d,s}(t) + u_{b,C}^{(i)}(t)\}$ (right panel) for successful identification ($d = s$) for two different participants - one performing well (dotted line) and one performing poorly (dashed line). The shaded areas represent the corresponding 90% point wise credible intervals. Parameters for the high-level tone response category T1 are shown in red; low-rising T2 in blue; low-dipping T3 in green; and high-falling T4 in purple.

by-trial variability is explained by stochastically different diffusion paths, the LBA model explains trial-by-trial variability assuming the slopes μ for different trials to be drawn from a Normal$(m_{d,s}, v_{d,s})$ distribution.

The literature on LBA models has many serious limitations. The normality assumption on the slopes μ clearly does not satisfy any non-negativity constraints. Existing LBA models are also limited in their use of a common boundary b_s for all decision categories d. There is also no principled way to incorporate systematic stimulus and decision category specific fixed or individual specific random effects into the LBA model. Existing literature is also limited to static settings, there is no mechanism to estimate smoothly varying longitudinal parameter trajectories as the participants get trained and experienced in their decision tasks. In our implementa-

tion, we thus fitted the LBA model separately for each block. Finally, the likelihood function of the LBA model is non-convex in the parameters. Parameter estimation based on optimization of the likelihood function is thus fraught with convergence issues. We used the `rtdists` package in `R`, using several random initializations and tracking the objective function to ensure convergence.

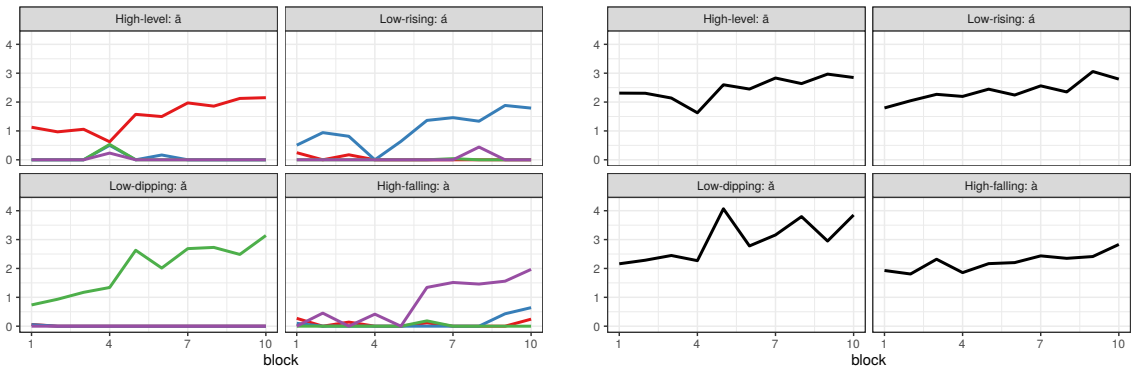


Figure 3.14: Results for tone learning data: Left: Estimated mean slopes $m_{d,s,t}$ for the LBA model. Right: Estimated boundaries $b_{s,t}$ for the LBA model. In the left panel, $m_{d,s,t}$'s for the high-level tone response category T1 are shown in red; low-rising T2 in blue; low-dipping T3 in green; and high-falling T4 in purple.

Results produced by the LBA model applied to our motivating tone-learning data are reported in Figure 3.14. Owing to the limitations discussed above, the inference we make with such models is very limited. For instance, only non-smooth population level estimates are available, individual specific trajectories can not be assessed, etc. Some of our findings can, however, be confirmed by the LBA method. For example, looking at the drift parameter estimates, one can see that tone T3 is consistently associated with larger drifts. As was also seen in the estimates returned by our method, tones {T2, T4} have similar values for the drift and the boundary

parameters. Except such general overall findings, the LBA model, however, can not answer scientific questions related to the dynamics of category learning with fine detail.

Our method, on the other hand, provides a biologically interpretable, statistically principled approach to accommodate fixed effects of input stimuli and decision categories as well as random subject specific heterogeneity, allows MCMC algorithm based efficient estimation of longitudinally smoothly evolving parameter trajectories, borrowing information across sample subgroups, participants as well as adjacent time stamps through many layers of hierarchy. Crucially, building on a novel local cluster inducing mechanism, our method also allows automated assessment of local similarities and differences in the parameter trajectories in very fine detail as the participants get trained and experienced in their decision tasks.

On the scientific side, the detailed insights obtained here point toward interesting and novel hypotheses about learning. For example, we demonstrate that a difference in drift rates, associated with the speed of sensory evidence accumulation, is critical in determining good vs poor learners. Evidence thresholds, on the other hand, remain relatively stable over training blocks as well as across participants. Recent studies have shown that the process of evidence accumulation can be selectively targeted by brain stimulation (Van der Groen *et al.*, 2018). Novel tone learning studies are currently being designed to test if such neurostimulation primarily improves the drift rates but not the evidence thresholds.

On the practical side, the insights obtained above can have important implications for developing advanced training regimens in language learning platforms

used by millions of adults. Due to poor understanding of the temporal dynamics of learning, especially in multi-category learning problems, current training regimens are neither time adaptive nor individualized. Similar to personalized medicine, next-generation speech training paradigms seek to optimize and individualize training to reduce vast inter-individual differences in learning success (Wong *et al.*, 2017; Birdsong, 2004). With our ability to assess detailed longitudinal confusion patterns, we can set up efficient training paradigms that can change the dynamics of learning in specific ways. For example, learners may generally benefit from introducing greater variability in pitch height that allows them to shift their focus on pitch direction and hence can reduce disparities in tone confusions like that between T2 and T4; poor learners may additionally benefit from 'perceptual fading' - beginning with easy tones like {T1,T3} and making the training more challenging afterward with the introduction of tones like {T2,T4}; etc. As mentioned before, non-invasive and safe brain stimulation approaches like transcranial random noise stimulation and vagus nerve stimulation can be leveraged to selectively improve the process of sensory accumulation that could enhance the performance in poor learners.

3.6 Linear Ballistic Accumulator Model

We present here a review of the LBA model (Brown and Heathcote, 2008) for easy reference.

The LBA model is a popular framework for studying neural mechanisms underlying choice between multiple alternatives. Similar to our model, it uses independent evidence accumulators starting at δ_s that continue until a response boundary

b_s is reached. The accumulator that first reaches the boundary corresponds to the decision outcome, and the time at which the boundary is reached is the response time. The evidence, however, accumulates linearly at the rate $\mu_{d,s}$, reaching the boundary b_s precisely at time $\tau_d = b_s/\mu_{d,s}$. To explain trial-by-trial variability, the LBA model assumes that the slopes μ for different trials are random draws from a Normal$(m_{d,s}, v_{d,s})$ distribution. The cumulative distribution function for the boundary crossing time τ_d for the d^{th} category is thus given by

$$F_{LBA}(\tau_d \mid \boldsymbol{\theta}_{d,s}) = 1 - \Phi\left(b_s/\tau_d \mid m_{d,s}, v_{d,s}\right),$$

where $\boldsymbol{\theta}_{d,s} = (m_{d,s}, v_{d,s}, b_s)^{\mathrm{T}}$. The likelihood of the LBA model at the t^{th} time point is thus

$$L_t(\mathbf{y}_t \mid \mathbf{s}, \boldsymbol{\theta}) = \prod_{d=1}^{d_0} \prod_{i=1}^{n} \prod_{\ell=1}^{L} \left[f_{LBA}(\tau_{i,\ell,t} \mid \boldsymbol{\theta}_{d,s,t}) \prod_{d' \neq d} \{1 - F_{LBA}(\tau_{i,\ell,t} \mid \boldsymbol{\theta}_{d',s,t})\} \right]^{\mathbb{1}\{d_{i,\ell,t}=d\}},$$

where $\boldsymbol{\theta}_{d,s,t} = (m_{d,s,t}, v_{d,s,t}, b_{s,t})^{\mathrm{T}}$, and $f_{LBA}(\tau) = \frac{dF_{LBA}(\tau)}{d\tau}$ is the pdf of τ.

The existing literature on LBA models has many serious limitations. The normality assumption on the slopes μ in the LBA model does not satisfy a non-negativity constraint. A common boundary b_s for all decision categories d is also inflexible. Importantly, there is no principled method to incorporate systematic stimulus and decision category specific fixed or individual specific random effects into the LBA model. Existing literature is also limited to static settings, there is no mechanism to estimate smoothly varying longitudinal trajectories as the participants get trained and experienced in their decision tasks. In our implementation, we thus fitted these models separately for each time stamp. Finally, the likelihood function

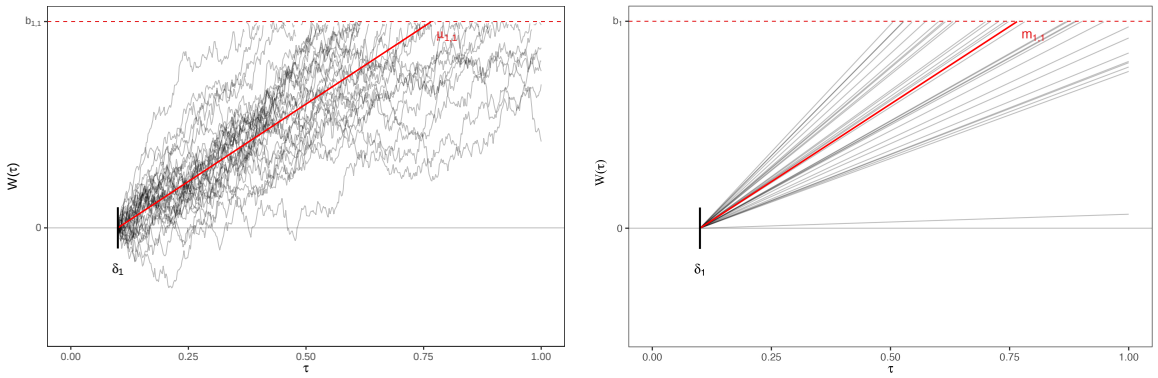


Figure 3.15: Representation of the underlying evidence accumulation processes for our drift-diffusion model (left) and the LBA model (right) for 30 independent trials with fixed stimulus and decision categories $d = s = 1$. The red line represents the drift parameter $\mu_{1,1}$ for the drift-diffusion model (left) and the mean of the drift parameters $m_{1,1}$ for the LBA (right). In drift-diffusion models, trial-by-trial variability is explained by stochastically different diffusion paths for different trials. In the LBA model, trial-by-trial variability is explained by stochastically varying slopes drawn from a Normal distribution.

of the LBA model described above is non-convex in the parameters. Parameter estimation based on optimization of the likelihood function is thus fraught with convergence issues. We used the `rtdists` package (Singmann *et al.*, 2019) in `R`, using several random initializations and tracking the objective function to ensure convergence.

3.7 Comparison with a Simpler Sub-Model

In this section, we summarize the results produced by a simpler alternative model, specifically, a reduced static version of our proposed longitudinal drift-diffusion mixed model fitted separately to data from each block as in the case

of the LBA model. Using notation similar to those in our proposed longitudinal mixed model, we now let $\mu_{x,t}^{(i)} = \exp\{f_{\mu,x,t} + u_{\mu,x,t}^{(i)}\}$ be the drift rates and $b_{x,t}^{(i)} = \exp\{f_{b,x,t} + u_{b,x,t}^{(i)}\}$ be the boundary parameters. The time index t now appears in subscript, as opposed to as an argument within parenthesis in our original longitudinal functional model. Other relevant parts of the model, including the priors, remain unchanged.

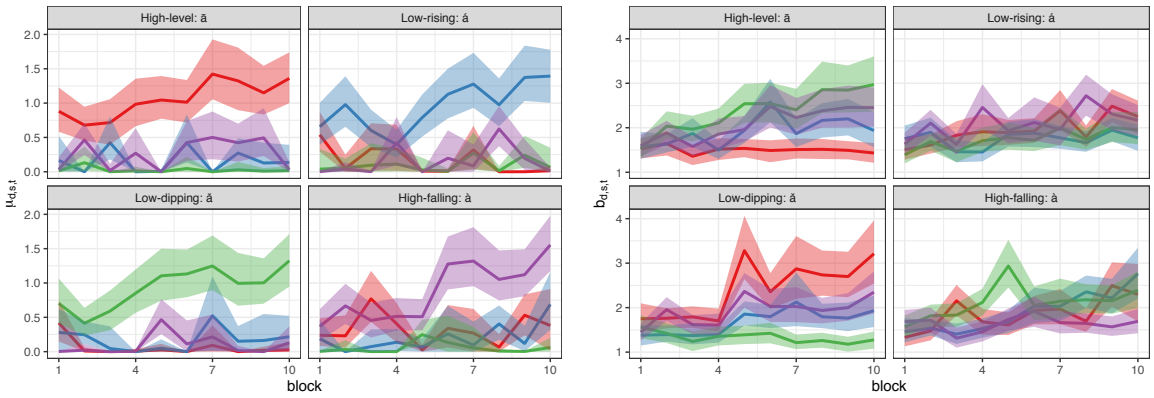


Figure 3.16: Results for tone learning data: Estimated posterior mean trajectories of the population level drifts $\mu_{d,s,t}$ (left panel) and boundaries $b_{d,s,t}$ (right panel) for the inverse Gaussian drift-diffusion mixed model applied independently for each block. The shaded areas represent the corresponding 90% point wise credible intervals. Parameters for the high-level tone response category T1 are shown in red; low-rising T2 in blue; low-dipping T3 in green; and high-falling T4 in purple.

Figure 3.16 shows the posterior means and associated 90% credible intervals for the population level boundaries $b_{d,s,t}$ and drift rates $\mu_{d,s,t}$ estimated by fitting the above described static drift-diffusion model fitted separately to data from each block. These results are generally consistent with the ones illustrated in Figure 3.11. However, this reduced model yields less interpretable results for at least three reasons. First, the absence of functional dependence makes it harder to pinpoint a general

trend because the estimates are not smooth but very wiggly across the training blocks. Second, the fixed effects parameters are not allowed to cluster across input-response combinations, which results in many redundant configurations. Third, the parameter estimates under our proposed model seem to have smaller uncertainty due borrowing of information across adjacent blocks as well as across input-output tone combinations via local clustering.

3.8 Simulation Studies

In this section, we discuss the results of some synthetic numerical experiments. We are not aware of any other method from the existing literature that can be readily applied or at least be easily adapted to our data settings and inferential challenges. We thus restrict our focus mostly on evaluating the performances of the proposed longitudinal inverse Gaussian drift-diffusion mixed model. We do present a comparison with the LBA model though, applying it separately for each block as in Section 3.5.

In designing the simulation scenarios, we have tried to closely mimic our motivating tone learning data set. We thus chose $n = 20$ participants being trained over $T = 10$ blocks to identify $d_0 = 4$ tones. We set $\mu_{d,s}(t), b_{d,s}(t)$ to values that are very similar to the corresponding estimated values for the real data set. The local differences were all set to be in the drift curves; additionally, some boundary trajectories were globally different from each other. We slightly simplified the local clustering structure, however, to be able to better illustrate the workings of our proposed method. Moreover, we choose $u_{\mu}^{(C,i)}(t), u_{b}^{(C,i)}(t), u_{\mu}^{(I,i)}(t), u_{b}^{(I,i)}(t), \delta_s$ etc. to

be the estimated posterior means obtained for the real data set.

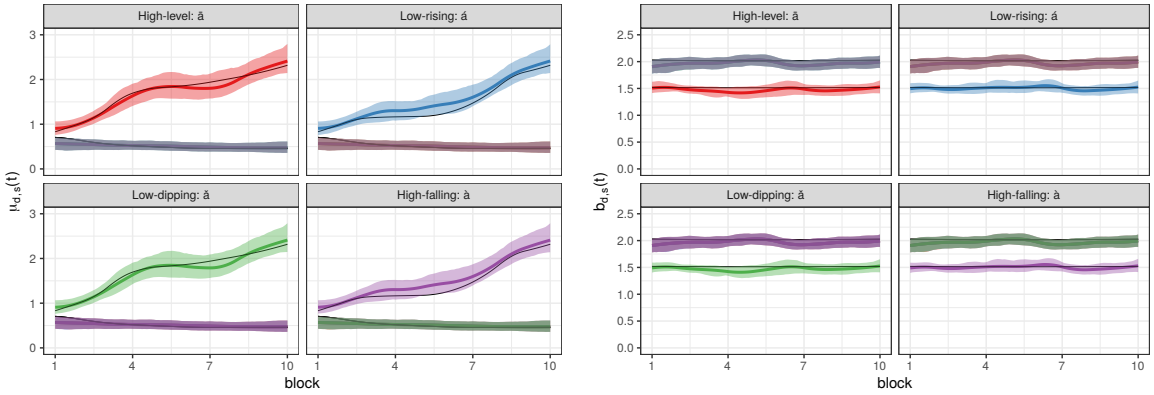


Figure 3.17: Results for synthetic data: Estimated posterior mean trajectories of the population level drifts $\mu_{d,s}(t)$ (left panel) and boundaries $b_{d,s}(t)$ (right panel) for the proposed longitudinal inverse Gaussian drift-diffusion mixed model. The shaded areas represent the corresponding 90% point wise credible intervals. The solid black lines represent underlying true curves. Parameters for the high-level tone response category T1 are shown in red; low-rising T2 in blue; low-dipping T3 in green; and high-falling T4 in purple.

We experimented with 50 synthetic data sets generated according to the design described above. The results produced by our method were highly stable and consistent across all data sets. The results summarized below represent a typical scenario.

Figure 3.17 shows the posterior mean trajectories and associated 90% credible intervals for the the drift rates $\mu_{d,s}(t)$ and boundaries $b_{d,s}(t)$, for every possible combination of (d, s). This figure suggests that the underlying true curves are all recovered well by our method. In comparison, the results obtained by the LBA model, displayed in Figure 3.18, suffer from the same limitations discussed in Section 3.5. Furthermore, Figures 3.19 and 3.20 suggest that the underlying true local partition

structure, as well as the individual specific parameter trajectories, are also estimated quite well by our method.

Figure 3.18 presents the results obtained by the LBA model applied to the synthetic data set. There is a general agreement between the population level estimates produced by our method and the LBA. However, as discussed in detail in Section 3.5 and Section 3.6, the LBA model has many serious limitations, including being incapable of producing individual level estimates, having shared boundary parameters across all input tones, not borrowing any information across adjacent time stamps etc. Only a very limited set of inferential questions can therefore be answered by the LBA model.

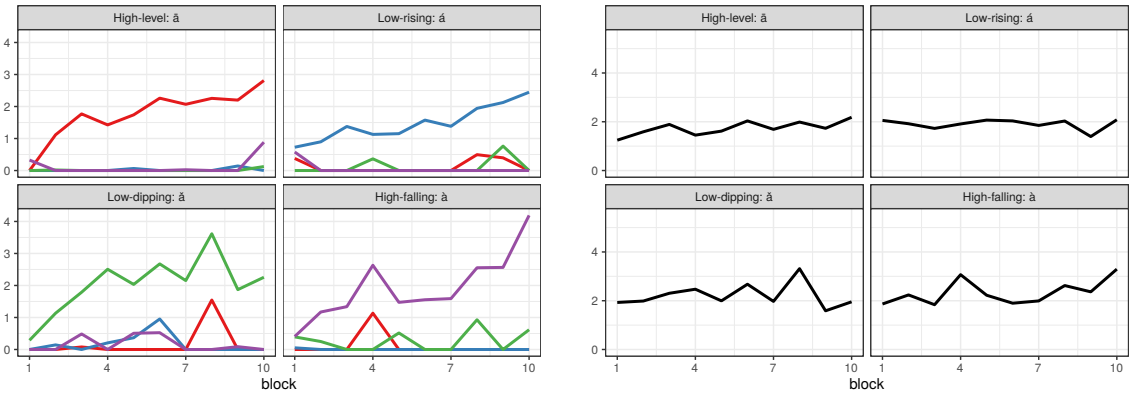


Figure 3.18: Results for synthetic data: Left: Estimated mean slopes $m_{d,s,t}$ for the LBA model. Right: Estimated boundaries $b_{s,t}$ for the LBA model. In the left panel, $m_{d,s,t}$'s for the high-level tone response category T1 are shown in red; low-rising T2 in blue; low-dipping T3 in green; and high-falling T4 in purple.

3.9 Discussion

Summary: In this chapter, we proposed a novel longitudinal drift-diffusion

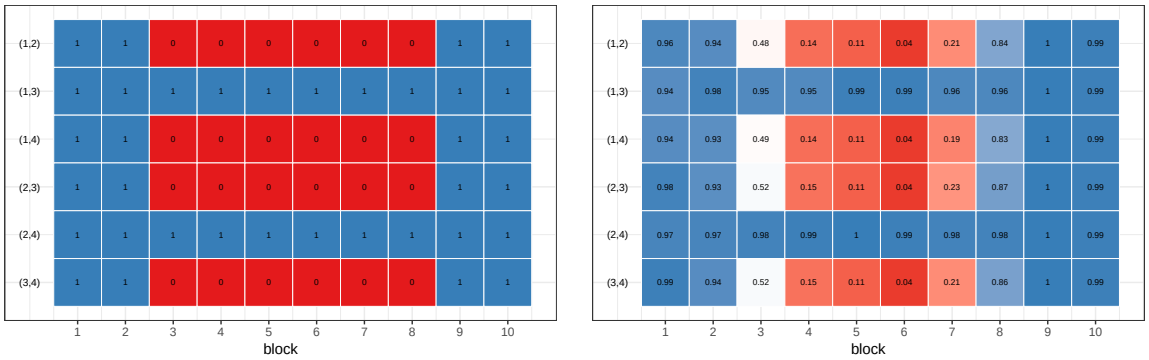


Figure 3.19: Results for synthetic data: The left panel shows the true clustering structure of the underlying parameter trajectories for successful identification ($d = s$) of different input tones in different learning phases. The right panel shows the corresponding posterior co-clustering probabilities estimated by our proposed method.

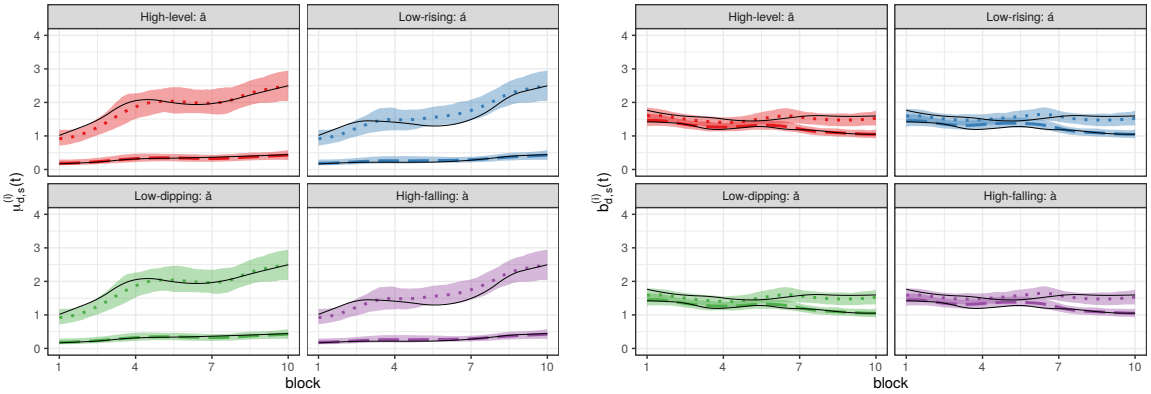


Figure 3.20: Results for synthetic data: Estimated posterior mean trajectories for individual specific drifts $\mu_{d,s}^{(i)}(t)$ (left panel) and boundaries $b_{d,s}^{(i)}(t)$ (right panel) for two different participants - one performing well (dotted line) and one performing poorly (dashed line). The shaded areas represent the corresponding 90% point wise credible intervals. The solid black lines represent underlying true curves. Parameters for the high-level tone response category T1 are shown in red; low-rising T2 in blue; low-dipping T3 in green; and high-falling T4 in purple.

mixed model for perceptual decision making, allowing the underlying mechanisms to be similar or different at different longitudinal stages. Our research was motivated primarily by auditory neuroscience experiments where scientists are interested in

understanding how the decision making mechanisms evolve as the participants get more training in the decision tasks. Our model was built on a novel statistical framework for longitudinal data that exploited local support properties of B-spline bases and (factorial) HMMs to allow automated assessment of local similarities and differences in the underlying parameter trajectories.

Application to our motivating tone categorization experiments provided interesting novel insights into the underlying learning mechanisms. Notably, we discovered that the improvements and the local variations in tone categorization performance can be explained mostly by variations in the underlying drift parameters while the boundaries mostly remain constant. We also discovered local groupings among the underlying parameter curves in various phases of the learning experiments, how they differ between well and poorly performing participants etc. Such inferences were outside the scope of the previously existing literature.

Methodological extensions: Methodological extensions and topics of our ongoing research include adapting the proposed models to time constrained learning experiments, developing nested models to capture the dynamics within the blocks, accommodating sleep induced overnight 'consolidation' effects, fully developing the inverse-probit model (3.4) for accuracies introduced in Section 3.3, etc.

Broader scientific impact: The proposed approach, we believe, takes the existing literature on drift-diffusion decision making models many significant steps forward, enabling neuroscientists to study the longitudinal behavior of biologically interpretable model parameters in much finer detail than what previous methods could achieve.

As reported in Section 3.5, the findings of our motivating speech learning experiment help formulate interesting novel scientific hypotheses about speech learning. The findings are also practically highly significant in providing exciting opportunities for developing time adaptive and individualized training regimens for language learning.

Efficient estimation of group and individual level trajectories also open exciting avenues for potential adaptations in clinical settings, especially in conjunction with simultaneously performed imaging studies.

Finally, the scope of proposed method is also not restricted to auditory neuroscience problems but the approach can be readily applied to study decision making mechanisms in other areas of neuroscience as well.

Chapter 4

Bayesian Semiparametric Longitudinal Functional Mixed Models with Locally Informative Predictors

We extend the methodology for locally varying partitions introduced in the previous chapter to the case of multiple predictors. In particular, we present a flexible Bayesian semiparametric mixed model for longitudinal functional data analysis in the presence of potentially high-dimensional categorical covariates. Our proposed method allows the fixed effects components to vary between dependent random partitions of the covariate space at different time points. The mechanism not only allows different sets of covariates to be included in the model at different time points but also allows the selected predictors' influences to vary flexibly over time. Smooth time-varying additive random effects are used to capture subject specific heterogeneity. We establish posterior convergence guarantees for both function estimation and variable selection. We design a Markov chain Monte Carlo algorithm for posterior computation. We evaluate the method's empirical performances through synthetic experiments and demonstrate its practical utility through real world applications.

4.1 Introduction

We propose a novel statistical framework for modeling longitudinally varying continuous response trajectories in the presence of categorical covariates. Such settings may be viewed as longitudinal adaptations of static analysis of variance (ANOVA) designs and hence are very generic and almost ubiquitously encountered in modern scientific research in many diverse fields, examples from recent statistics literature including pharmacodynamics (De Iorio *et al.*, 2004), mass spectroscopy (Morris and Carroll, 2006), early pregnancy loss studies (MacLehose and Dunson, 2009), etc. In such scenarios, assessing the local variations in the response profiles, including especially how the associated predictors might influence the response differently in different stages of the longitudinal process, can provide valuable insights into the underlying data generating mechanisms.

Existing Methods: The literature on longitudinal data analysis is really vast (see, for example, books by Diggle *et al.*, 2002; Singer *et al.*, 2003; Fitzmaurice *et al.*, 2008, and the references therein). Bayesian methods for longitudinal data have also been extensively developed (Daniels and Pourahmadi, 2002; Chib and Hamilton, 2002; Li *et al.*, 2010; Müller *et al.*, 2013; Quintana *et al.*, 2016, etc.). However, the problem of characterizing dynamically varying variable importance in such settings has not received much attention. This chapter presents a novel Bayesian semiparametric method that addresses such needs.

Our work in this direction was inspired by the existing sparse literature on local clustering in functional data (Duan *et al.*, 2007; Petrone *et al.*, 2009; Nguyen, 2010; Nguyen and Gelfand, 2011). These Bayesian nonparametric approaches assume

that the mean functions of interest can be represented by a smaller set of canonical curves that are in turn modeled, for instance, as independently and identically distributed (i.i.d.) realizations from a stationary Gaussian process. Gelfand *et al.* (2005) specify an infinite mixture of these global functional atoms in which each observation is a noisy realization around a draw from the set of canonical curves. Such an approach allows for curves that are either completely different or completely identical across the entire functional domain, capturing only their global difference patterns. Many applications, however, involve data exhibiting local heterogeneity. Local clustering in such cases could greatly improve estimation and prediction, borrowing information across locally homogenous regions, as well as interpretability and inference, providing potentially interesting insights into the underlying causes of local heterogeneity. Toward this goal, Duan *et al.* (2007) proposed a solution by defining a stick-breaking construction at each location, which allows for local selection of curves. Petrone *et al.* (2009) assumed that the individual curves can be obtained as hybrid species defined as recombinations of different portions of the canonical curves. Both these approaches define the local allocation rules by means of a single hidden labeling process that indicates which canonical curve is chosen at each time stamp. Additional challenges are represented by the choice of functional dependence in the labeling process, whose theoretical properties have been studied by Nguyen and Gelfand (2011). Suarez and Ghosal (2016) proposed an alternative approach, using independent priors at different time points to cluster wavelet basis coefficients first, but then using these local features to find global functional clusters as the final inference goal.

The approaches mentioned above have limitations that deserve attention. First, defining the mean functions as recombinations of canonical curves implies that these curves are discontinuous, which can be an impractical assumption in most applications. Continuous curves may be desired, for instance, in dose-response relationships (De Iorio *et al.*, 2004). Second, the inclusion of covariates in these models has only been accomplished via an additive term in the mean function. Even when flexible random effects are used, the linearity assumption of the covariate effects can be quite restrictive in practice. Furthermore, the problem of dynamically characterizing variable importance in these settings has not been addressed.

Alternative approaches to model time-varying predictor effects and interactions in longitudinal data include tree based methods. Bayesian additive regression trees (BART) (Chipman *et al.*, 2010) perform well when the regression function consists of low order nonlinear interactions. With time as an additional covariate, these models can be adapted to capture longitudinally varying influences of the predictors (Sparapani *et al.*, 2016). Separate ideas involving a single tree have also been proposed (Taddy *et al.*, 2011; Gramacy *et al.*, 2013) where the tree structure evolves when new data streams become available. Linero and Yang (2018) and Starling *et al.* (2020) proposed smoothing the covariate effects which yields more appropriate results when the outcome is expected to vary smoothly over time. These models, albeit flexible, do not directly assess the local influence of each individual predictor but measure variable importance by calculating their contributions to reducing the in-sample mean squared error. With such heavy emphasis on prediction, they often include many weakly informative or even spurious predictors in the ensembles and

lack parsimony and interpretability as a result (Breiman, 2001; Efron, 2020).

Yet another related strategy comprises varying coefficients (VC) regression models where the regression coefficients are allowed to smoothly vary over a set of chosen modifiers (Hastie and Tibshirani, 1993). VC models have been adapted to longitudinal data by considering time as the only modifier (Hoover *et al.*, 1998). More recently BART priors (Deshpande *et al.*, 2020) and variable selection techniques (Koslovsky *et al.*, 2020) have also been adapted to VC settings. While VC models allow for an easy assessment of the predictors' importance, they are restricted in their ability to accommodate interactions between predictors. For example, for p categorical predictors $x_j \in \{1, \ldots, x_{j,\max}\}, j = 1, \ldots, p$, it is necessary to include $\sum_{j=1}^{p}(x_{j,\max}-1)$ dummy variables for the main effects, $\sum_{j_1 \neq j_2}(x_{j_1,\max}-1)(x_{j_2,\max}-1)$ for the first order interactions, and so on.

Our Proposed Approach: We propose a longitudinal functional mixed effects model that combines predictive power and interpretability by addressing the limitations of the local clustering approaches cited above. Most existing methods imply a tension between the main goals of statistical analysis (Breiman, 2001), namely estimation, attribution and prediction (Efron, 2020). Our proposed approach tries to strike a balance - it is highly flexible, being able to accommodate higher order interactions between the predictors, but also favors parsimony, modeling these complex effects implicitly and compactly, while also allowing some ease of interpretation, including explicitly encoding each predictor's varying overall significance at different time points. Our method also comes with theoretical guarantees for both function estimation and variable selection. Table 4.1 shows what different classes of methods

can achieve when applied to longitudinal data with locally informative predictors.

Method	Fixed effects	Random effects	Variable selection	Dynamic partitions	Complex interactions	Easy interpretation
Our proposed LFMM	✓	✓	✓	✓	✓	✓
BART methods	✓	×	✓	×	✓	×
VC methods	✓	×	✓	×	×	✓
LASSO methods	✓	✓	✓	×	×	✓

Table 4.1: A broad comparison of what different existing approaches are able to infer in regression for longitudinal data with categorical predictors. No other method allows for borrowing of information across covariate levels through dynamic partitions.

The construction of our proposed model proceeds by characterizing the longitudinal evolution of both the predictor dependent fixed effects and the subject specific random effects as flexible functions of time (Ramsay and Silverman, 2007; Morris, 2015; Wang *et al.*, 2016) modeled by mixtures of locally supported spline bases (de Boor, 1978; Eilers and Marx, 1996). The fixed effects model spline coefficients are allowed to vary with the associated predictors' level combinations, thereby accommodating all order interactions between them. Structuring these coefficients as multi-way tensors and applying a novel higher order singular value (HOSVD) type decomposition (Tucker, 1966; De Lathauwer *et al.*, 2000; Kolda and Bader, 2009), we reduce the high-dimensional problem of modeling the complex joint influence of many different predictors to that of estimating much smaller-dimensional core coefficients. In effect, this induces a local partitioning of the joint covariate space such that the different predictor level combinations belonging to the same partition set will have a similar effect on the response variable. The local partitions constructed this way can in fact be indexed by combinations of separate latent allocation indicators, one for

each level of the associated categorical predictors, facilitating separate assessment of the influences of each individual covariate (Sarkar and Dunson, 2016). To induce dependence between the adjacent local partitions, we allow the latent allocation indicators evolve according to a factorial hidden Markov model (fHMM) (Ghahramani and Jordan, 1997). In constructing the model this way, we break free from the assumption of separate canonical curves of the previously existing Bayesian nonparametric literature cited above but allow the dependencies across adjacent temporal locations be further informed by the associated local partition configurations through a novel conditionally Markov prior on the core spline coefficients, conditional on the partition structure, improving model interpretability and estimation efficiency. The proposed functional approach also has the important advantage of avoiding to have to impute missing data when they are missing under simple mechanisms (Little and Rubin, 2019). We establish theoretical results on posterior consistency of the proposed method for both function estimation and variable selection. We evaluate its numerical performance in simulation experiments where it significantly outperformed its competitors not just on average but also uniformly in all simulation instances. Finally, we illustrate the method's practical performance in real data applications from diverse domains.

The methodology presented here is highly generic and broadly adaptable to diverse problems. For instance, Paulon *et al.* (2020) developed a similar local clustering method in the presence of a single categorical predictor x with a small number of levels for a specific application with a complex drift-diffusion likelihood function. The focus of this work, however, is on developing a general methodology with an emphasis

on the multivariate case $(x_1, \ldots, x_p)$ which presents significant and unique additional modeling and computational challenges. For instance, redefining the $\prod_{j=1}^{p} x_{j,\max}$ level combinations of $(x_1, \ldots, x_p)$ as the levels of a new single predictor x, while conceptually straightforward, does not provide a practically effective solution as it does not allow separate characterization of the local importances of the different predictors and, with $\prod_{j=1}^{p} x_{j,\max}$ increasing exponentially fast with p, quickly becomes computationally inefficient even in small to moderate dimensional problems. The strategy is practically useless, for instance, in a real data applications we discuss in Section 4.6, where $\prod_{j=1}^{p} x_{j,\max} = 580,608$. Our proposed dynamic HOSVD based approach, in contrast, not only provides a flexible and highly efficient tool for dimension reduction and simultaneous variable selection but also does this locally at each time point while borrowing information across a number of levels.

Our proposed approach to flexible longitudinal mixed model regression and simultaneous variable selection does not partition the response values directly, which has been considered by many in the static setting (Hartigan, 1990; Denison *et al.*, 2002; Quintana and Iglesias, 2003) and by some in the dynamic setting (Barry and Hartigan, 1992; Page *et al.*, 2020). Instead, we partition the covariate space according to their influences on the response. Separately, the literature on HMMs and fHMMs is also vast (Rabiner, 1989; Scott, 2002; Frühwirth-Schnatter, 2006; Zucchini *et al.*, 2017). To our knowledge, however, they have never been adapted in the novel ways proposed in this chapter to dynamic variable selection problems. There is also a growing body of literature on regression methods for tensor valued predictors with tensor factorization techniques used as a dimension reduction tool. These methods,

however, apply tensor factorizations with all continuous components, where the general Tucker decomposition runs into identifiability and interpretability problems. To avoid these issues, the literature has focused on parallel factor (PARAFAC) type decomposition (see, e.g., Guhaniyogi *et al.*, 2017; Papadogeorgou *et al.*, 2019, etc.), a much simpler but restrictive special case of the Tucker. Aside from the development of sophisticated dependence models for the tensor components in a longitudinal setting, our proposal is also novel in that we employ a compact HOSVD, a flexible but interpretable version of the Tucker decomposition, where the core tensors take continuous values but the mode matrices comprise specially structured binary elements, resulting in interpretable partition structures that allow dynamic variable selection.

Outline of the Chapter: The rest of this chapter is organized as follows. Section 4.2 develops the generic longitudinal mixed model framework. Section 4.3 establishes posterior convergence guarantees for the proposed model, for both function estimation and variable selection. Section 4.4 develops Markov chain Monte Carlo (MCMC) algorithms for posterior computation. Section 4.5 presents the results of simulation experiments. Section 4.6 presents real data applications. Section 4.7 contains concluding remarks.

4.2 Longitudinal Functional Mixed Model

In this section, we develop a novel generic statistical framework for longitudinal functional mixed model (LFMM), where a response y is generated under the influence of p categorical predictors $x_j \in \{1, \ldots, x_{j,\max}\} = \mathcal{X}_j, j = 1, \ldots, p$ longitudinally over time. To be precise, data y_{i,ℓ,t_i}, available for individuals $i \in \{1, \ldots, n\}$ and

trials $\ell \in \{1, \ldots, L_{i,t_i}\}$ at time points $t_i \in \{t_{i,1}, \ldots, t_{i,T}\}$, are generated under the influence of the predictors $x_j, j = 1, \ldots, p$. Importantly, we are not only interested in assessing the overall global influences of the predictors but also how they affect the responses locally at various times of the longitudinal studies.

We consider the following generic class of LFMMs

$$\{y_{i,\ell,t} \mid x_{j,i,\ell,t} = x_j, j = 1, \ldots, p\} = f_{x_1,\ldots,x_p}(t) + u_i(t) + \varepsilon_{i,\ell,t}, \qquad \varepsilon_{i,\ell,t} \sim f_\varepsilon, \quad (4.1)$$

where $f_{x_1,\ldots,x_p}(t)$ denotes time-varying fixed effects due to associated predictors $\mathbf{x} = (x_1, \ldots, x_p) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_p = \mathcal{X}$, $u_i(t)$ are time-varying subject specific random effects, and $\varepsilon_{i,\ell,t}$ are random errors, i.i.d. from f_ε, satisfying $\mathbb{E}_{f_\varepsilon}(\varepsilon_{i,\ell,t}) = 0$. We assume that $f_{x_1,\ldots,x_p}(t)$ and $u_i(t)$ evolve continuously with time. In this work, we focus on normally distributed errors with an inverse-Gamma prior on the error variance as

$$f_\varepsilon = \text{Normal}(0, \sigma_\varepsilon^2), \quad \sigma_\varepsilon^2 \sim \text{Inv-Ga}(a_\sigma, b_\sigma).$$

For ease of exposition, we assume in (4.1) and henceforth that the data points are measured at a common set of equidistant time points $\{t_1, \ldots, t_T\}$, denoted simply as $\{1, \ldots, T\}$. With some abuse of notation, generic values taken by the response y, the predictors x_j are also denoted by y, x_j, etc. Without loss of generality, we also assume henceforth the same number of replicates $L_{i,t} = L$ for all i, t. To further simplify notation, generic data recording time stamps in $\{1, \ldots, T\}$ as well as other generic time points in $[1, T]$ will both be denoted by t.

For longitudinal data observed on a regular time grid, as in the setting considered in this work, continuous functional parameter trajectories may still be more

appealing and interpretable to a practitioner. A functional approach to modeling longitudinal data also does not require to impute missing data when they are missing at random (Little and Rubin, 2019).

The focus of this work is on continuous responses with categorical predictors. In many applications, the covariates are exogenous, that is, for each i, the $x_{j,i,\ell,t}$'s equal some fixed level x_j for all ℓ, t. When they are time-varying, we assume that all levels of x_j are present in the sample at each t for each j. A nonideal but easy and practically useful approach to include continuous and ordinal predictors in model (4.1) would be to categorize them by binning their values into intervals (for example, using their quantiles) or by ignoring their order. Non-continuous responses of various types can likewise be conveniently analyzed via latent continuous variable augmentations (Albert and Chib, 1993; Dunson, 2000; Polson *et al.*, 2013).

4.2.1 Fixed Effects Model

We propose a novel approach to model the latent functions $f_{x_1,\ldots,x_p}(t)$ using basis decomposition methods that allow them to flexibly vary with time t while also locally depend on the predictor combinations $(x_1, \ldots, x_p)$. Specifically, we let

$$f_{x_1,\ldots,x_p}(t) = \textstyle\sum_{k=1}^{K} \beta_{k,x_1,\ldots,x_p} B_k(t), \tag{4.2}$$

where $\mathbf{B}(t) = \{B_1(t), \ldots, B_K(t)\}^{\mathrm{T}}$ are a set of known locally supported basis functions and $\boldsymbol{\beta}_{x_1,\ldots,x_p} = \{\beta_{1,x_1,\ldots,x_p}, \ldots, \beta_{K,x_1,\ldots,x_p}\}$ are unknown coefficients to be estimated from the data. We use B-spline bases (de Boor, 1978) which are nonnegative, continuous and have desirable local support properties (Figure 2.1). Allowing the

$\beta_{k,x_1,\ldots,x_p}$'s to vary with all predictor combinations $(x_1,\ldots,x_p)$, the model also accommodates all order interactions among the predictors. In the following, we use knots at the observed locations, hence $K = T$. This choice favors model interpretability since knot locations represent the set of time points where local variable selection is performed. As noted in Ruppert (2002), once smoothing is controlled by a penalty parameter, the number of knots K is not a crucial parameter as long as it is larger than a minimum threshold.

For most practical applications, the size $K\prod_{j=1}^{p} x_{j,\max}$ of the unstructured model (4.2) may, however, be too big to allow efficient estimation of the parameters. It is also difficult to assess local influences of the predictors using such unstructured models. A potentially efficient solution that can greatly reduce dimensions while also facilitating the assessment of predictors' importance is to cluster the parameters by allowing them to have common shared values across different predictor combinations. If, for example, $\boldsymbol{\beta}_{x_1,\ldots,x_{j,1},\ldots,x_p} = \boldsymbol{\beta}_{x_1,\ldots,x_{j,2},\ldots,x_p}$ for all combinations of $(x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_p)$, then not only have we reduced the number of parameters to be modeled but have also established that the two levels $x_{j,1}$ and $x_{j,2}$ of x_j have no differential effect on the data generating mechanism.

Local Clustering via Multi-layer Partitions: Such global clustering of all elements of $\boldsymbol{\beta}_{x_1,\ldots,x_p}$ together will still be highly restrictive in most practical applications. More realistically, the elements of $\boldsymbol{\beta}_{x_1,\ldots,x_p}$ should be allowed to cluster locally. To achieve this, we introduce local random partitions $\boldsymbol{\rho}_k = \{\boldsymbol{\rho}_{k,1},\ldots,\boldsymbol{\rho}_{k,m_k}\}$ of $\mathcal{X}$, where m_k denotes the cardinality of $\boldsymbol{\rho}_k$, and let

$$\{f_{x_1,\ldots,x_p}(t) \mid (x_1,\ldots,x_p) \in \boldsymbol{\rho}_{k,h_k}, k = 1,\ldots,K\} = \textstyle\sum_{k=1}^{K} \beta^{\star\star}_{k,h_k} B_k(t), \tag{4.3}$$

or equivalently,

$$\{\beta_{k,x_1,\ldots,x_p} \mid (x_1,\ldots,x_p) \in \boldsymbol{\rho}_{k,h_k}, k = 1,\ldots,K\} = \beta^{\star\star}_{k,h_k}.$$

While other higher order B-splines can also be used, in this work, we use linear B-splines (Figure 2.1) whose local support properties result in locally linear approximations of the fixed effects function with a simple expression $\{f_{x_1,\ldots,x_p}(k) \mid (x_1,\ldots,x_p) \in \boldsymbol{\rho}_{k,h}\} = \beta^{\star\star}_{k,h}$ at the k^{th} knot point, allowing simpler interpretations for the local partitions $\boldsymbol{\rho}_k$'s and also easier theoretical treatment and posterior computation.

In applications with large p, defining local partitions of the predictors' space presents daunting computational challenges due to the possible inclusion of all p covariates. Such curse of dimensionality can be overcome by defining the local random partitions in two steps (see Figure 4.1). First, we perform local variable selection by clustering the levels of each of the p marginal predictors individually. Defining the joint local partition as the product of the p marginal partitions can be restrictive in practice as it often yields overparametrized models. Thus, in the second step we construct the final joint partition by further clustering the product of the marginals. The multi-layer nature of the proposed local random partitions allows for a parsimonious representation of the set of B-spline coefficients, as well as for an intuitive specification of the dynamic partitions for the marginal predictors.

To construct the first layer of the local random partitions, we introduce the latent variables $z^{(x_j)}_{j,k} \in \mathcal{X}_j$. When $z^{(x_{j,1})}_{j,k} = z^{(x_{j,2})}_{j,k} = z_{j,k}$ for two different levels $x_{j,1}$ and $x_{j,2}$ of the j^{th} predictor x_j, the spline coefficients at location k do not differ

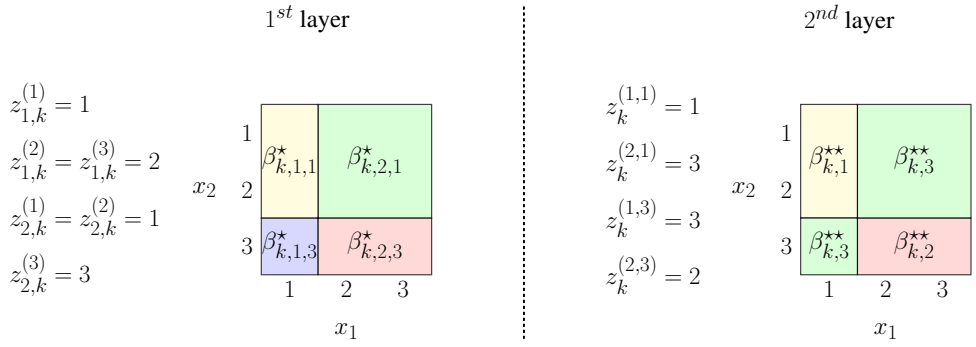


Figure 4.1: Illustration of the two layers of latent variables that induce the partition of the covariate space at a fixed location k in the case with two categorical predictors x_1, x_2 with $x_{1,\max} = x_{2,\max} = 3$ levels each. In this example, $\boldsymbol{\rho}_{k,1} = \{(1,1),(1,2)\}$, $\boldsymbol{\rho}_{k,2} = \{(2,3),(3,3)\}$, $\boldsymbol{\rho}_{k,3} = \{(1,3),(2,1),(2,2),(3,1),(3,2)\}$, $\ell_k = 4$, $m_k = 3$.

between $x_{j,1}$ and $x_{j,2}$, i.e. $\beta_{k,x_1,\ldots,x_{j,1},\ldots,x_p} = \beta_{k,x_1,\ldots,x_{j,2},\ldots,x_p} = \beta^{\star}_{k,z_{1,k},\ldots,z_{j,k},\ldots,z_{p,k}}$. There is thus no significant difference between how the two levels $x_{j,1}$ and $x_{j,2}$ influence the response y at location k. Letting

$$\mathcal{Z}_{j,k} = \{z_{j,k} : z^{(x_j)}_{j,k} = z_{j,k} \text{ for some } x_j \in \mathcal{X}_j\},$$
$$\mathcal{Z}_k = \{(z_{1,k},\ldots,z_{p,k}) : (z^{(x_1)}_{1,k},\ldots,z^{(x_p)}_{p,k}) = (z_{1,k},\ldots,z_{p,k}) \text{ for some } (x_1,\ldots,x_p) \in \mathcal{X}\},$$

the set of B-spline coefficients to be estimated at location k is at most $|\mathcal{Z}_k|$. When the $z^{(x_j)}_{j,k}$'s are assigned probability models supported on $\mathcal{X}_j$, the number of distinct values taken on by the $z^{(x_j)}_{j,k}$'s, namely $|\mathcal{Z}_{j,k}| = \ell_{j,k}$, may be less than $|\mathcal{X}_j| = x_{j,\max}$. The maximum size of the coefficient space $|\mathcal{Z}_k|$ at location k is $|\mathcal{Z}_k| = \prod_{j=1}^{p} \ell_{j,k} \leq \prod_{j=1}^{p} x_{j,\max} = |\mathcal{X}|$. Significant reduction in dimension is thus achieved at location k when $\prod_{j=1}^{p} \ell_{j,k} \ll \prod_{j=1}^{p} x_{j,\max}$. Importantly, the case $|\mathcal{Z}_{j,k}| = 1$ characterizes the scenario when x_j has no influence on y at location k. On the other end, when $|\mathcal{Z}_{j,k}| = x_{j,\max}$, the $z^{(x_j)}_{j,k}$'s take on different values for different levels of x_j, implying

that the spline coefficients are all different for different level of x_j at location k. All levels of x_j now significantly differently influence the response generating mechanism at location k.

Furthermore, we introduce the second layer of latent variables $z_k^{(z_{1,k},\ldots,z_{p,k})} \in \{1,\ldots,\ell_k\}$ with $\ell_k = \prod_{j=1}^p \ell_{j,k}$ that construct the final joint partition by further clustering the product marginals (Figure 4.1). Specifically, we let

$$\{\beta^\star_{k,z_{1,k},\ldots,z_{p,k}} \mid (z_{1,k}^{(x_1)},\ldots,z_{p,k}^{(x_p)}) = (z_{1,k},\ldots,z_{p,k}), z_k^{(z_{1,k},\ldots,z_{p,k})} = z_k\} = \beta^{\star\star}_{k,z_k}.$$

Dynamically Evolving Partition Structures: We now consider the problem of specifying probability models for the $z_{j,k}^{(x_j)}$'s that allow them to be temporally dependent across k. We model the temporal evolution of the $z_{j,k}^{(x_j)}$'s using hidden Markov models (HMMs). For each predictor combination $(x_1,\ldots,x_p)$, the collection $\mathbf{z}^{(x_1,\ldots,x_p)} = \{z_{j,k}^{(x_j)}, k = 1,\ldots,K, j = 1,\ldots,p\}$ then defines a factorial HMM (Ghahramani and Jordan, 1997) (see Section 2.6). We characterize the dynamics of the fHMM component chains as

$$(z_{j,k}^{(x_j)} \mid z_{j,k-1}^{(x_j)} = z_{k-1}) \sim \text{Mult}(\pi^{(j)}_{z_{k-1},1},\ldots,\pi^{(j)}_{z_{k-1},z_{j,\max}}) \quad \text{for } j = 1,\ldots,p.$$

We assign Dirichlet priors on the transition probabilities

$$\boldsymbol{\pi}_z^{(j)} = (\pi^{(j)}_{z,1},\ldots,\pi^{(j)}_{z,z_{j,\max}})^{\mathrm{T}} \sim \text{Dir}(\alpha^{(j)}/z_{j,\max},\ldots,\alpha^{(j)}/z_{j,\max}) \text{ with } \alpha^{(j)} \sim \text{Ga}(a_\alpha, b_\alpha).$$

In general, the maximum number of distinct values of the $z_{j,k}^{(x_j)}$'s is $x_{j,\max}$. However, in most applications, $|\mathcal{Z}_{j,k}|$ will be much smaller than $x_{j,\max}$ uniformly for all k

and the restricted support $z_{j,k}^{(x_j)} \in \{1, \ldots, z_{j,\max}\}$, $z_{j,\max} < x_{j,\max}$ will suffice. We impose parsimony by assigning exponentially decaying prior on the partition sizes $|\mathcal{Z}_{j,k}| = \ell_{j,k}$, favoring smaller partitions as

$$\ell_{j,k} \propto \exp(-\varphi_j \ell_{j,k}), \quad \varphi_j \sim \text{Ga}(a_{\varphi,j}, b_{\varphi,j}), \quad j = 1, \ldots, p, \quad k = 1, \ldots, K.$$

Larger values of φ_j here induce faster decay. Gamma hyper-priors on the φ_j's further make these shrinkage strengths data adaptive.

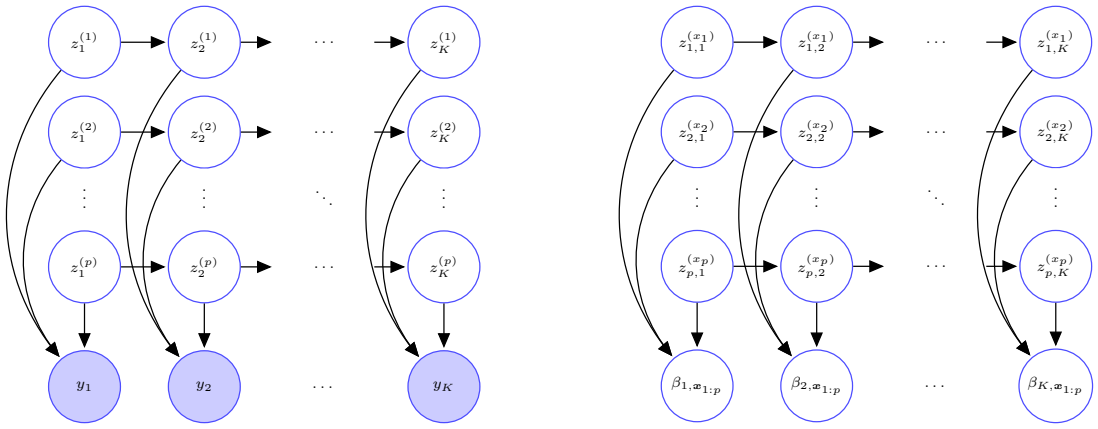


Figure 4.2: Left panel: The directed acyclic graph (DAG) of a conventional fHMM with p latent layers. Right panel: DAG of our proposed fixed effects model (4.3) with p categorical predictors $\mathbf{x}_{1:p} = (x_1, \ldots, x_p)$.

The second layer latent allocation variables $z_k^{(z_{1,k},\ldots,z_{p,k})}$ are assigned multinomial distributions with Dirichlet priors on the probability parameters as

$$(z_k^{(z_{1,k},\ldots,z_{p,k})} \mid \boldsymbol{\pi}_k^\star) \sim \text{Mult}(\pi_{k,1}^\star, \ldots, \pi_{k,\ell_k}^\star),$$
$$\boldsymbol{\pi}_k^\star = (\pi_{k,1}^\star, \ldots, \pi_{k,\ell_k}^\star)^{\mathrm{T}} \sim \text{Dir}(\alpha^\star/\ell_k, \ldots, \alpha^\star/\ell_k) \quad \text{with} \quad \alpha^\star \sim \text{Ga}(a_{\alpha^\star}, b_{\alpha^\star}).$$

When the $z_{j,k}^{(x_j)}$'s corresponding to two different categories of x_j are equal in a temporal region, the local support properties of B-splines then cause the underlying

curves to be the same in that region. Conversely, if the $z_{j,k}^{(x_j)}$'s corresponding to two different values of x_j are different, the underlying curves will be distinct unless the second layer of latent variables maps them to the same joint partition element.

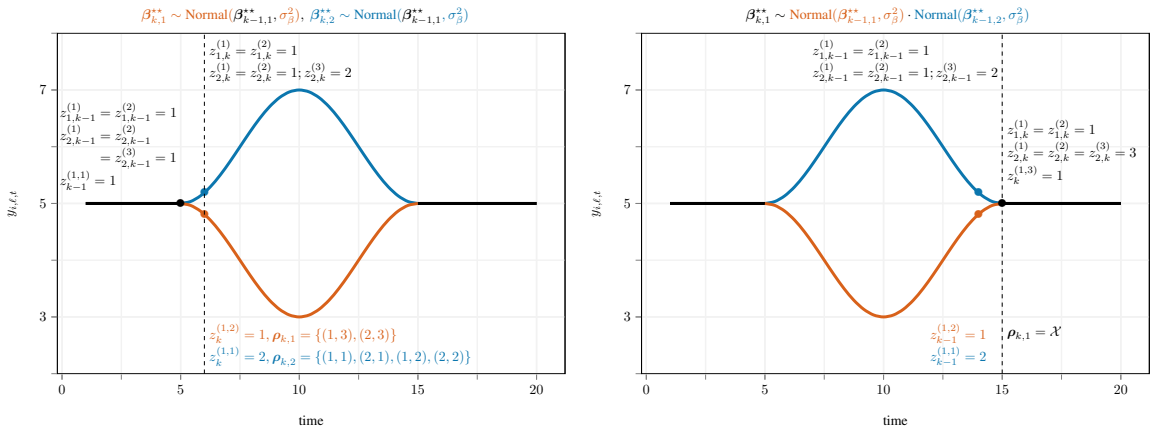


Figure 4.3: An illustration of the prior on the spline core coefficients $\beta_{k,z_k}^{\star\star}$ at location k (the dashed vertical lines) in the fixed effects model developed in Section 4.2.1 for a scenario with two categorical covariates $x_1 \in \{1,2\}$ and $x_2 \in \{1,2,3\}$, where the curves corresponding to all levels of (x_1, x_2) are initially equal, the curves for $x_2 = 1, 2$ (in blue) and $x_2 = 3$ (in red) then diverge at $t = 5$, merging back again at $t = 15$.

Conditionally Markov Regression Coefficients: We next consider priors for the atoms $\beta_{k,z_k}^{\star\star}$. Conditional on the $z_{j,k}^{(x_j)}$'s, $z_k^{(z_{1,k},\ldots,z_{p,k})}$'s, and the coefficients at the previous locations, for $k = 2, \ldots, K$, we construct the priors sequentially as

$$\beta_{k,z_k}^{\star\star} \sim \prod_{h \in \mathcal{Z}_{k,z_k}^{-}} \text{Normal}\left(\beta_{k-1,h}^{\star\star}, \sigma_\beta^2\right), \tag{4.4}$$

where $\mathcal{Z}_{k,z_k}^{-} = \left\{ z_{k-1} : z_{k-1} = z_{k-1}^{(z_{1,k-1},\ldots,z_{p,k-1})}; (z_{1,k-1}, \ldots, z_{p,k-1}) = (z_{1,k-1}^{(x_1)}, \ldots, z_{p,k-1}^{(x_p)}); (x_1, \ldots, x_p) \in \boldsymbol{\rho}_{k,z_k} \right\}$ and $\boldsymbol{\rho}_{k,z_k}$ is the partition element comprising the covariates levels $(x_1, \ldots, x_p)$ that, at location k, are assigned the label z_k. Simply put, we center

the core coefficients around the ones that are 'expressed' at the previous location (Figure 4.3), thus effectively penalizing their differences. The initial coefficients are assigned non-informative flat priors as $\beta^{\star\star}_{1,z_1} \sim 1$. The smoothness of the curves is thus controlled by the parameter σ^2_β and is assigned a prior, allowing it to be informed by the data. We let

$$\sigma_\beta \sim \mathrm{C}^+(0, s_\sigma),$$

where $\mathrm{C}^+(a, b)$ denotes a half-Cauchy distribution (Gelman, 2006; Polson and Scott, 2012) with location parameter a and scale parameter b.

In designing the model in this way, we deviate from existing approaches of informing the smoothness only through underlying canonical parameter curves (Petrone *et al.*, 2009; Nguyen and Gelfand, 2011, 2014), thereby improving model interpretation and estimation efficiency.

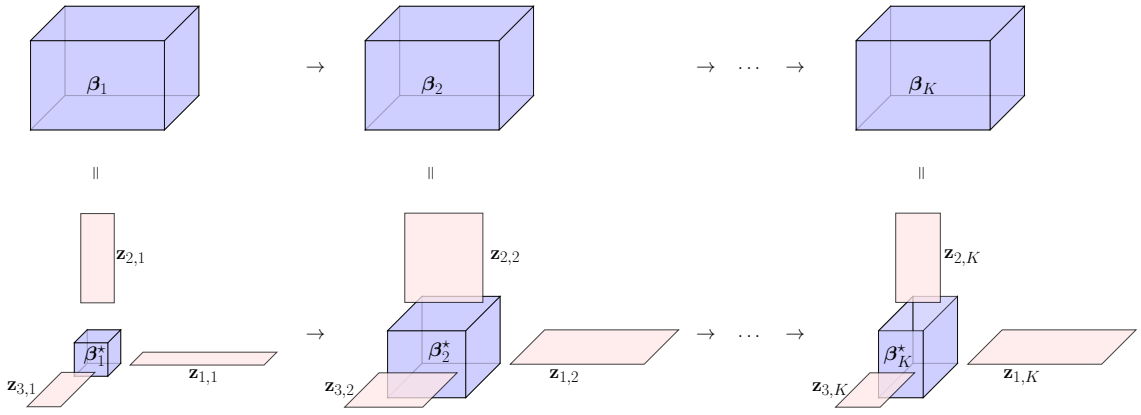


Figure 4.4: Model (4.3) with three covariates viewed as a dynamic HOSVD.

Alternative Characterization as HOSVD: The dimension reduction and local cluster inducing properties of our model can alternatively be understood and in

fact were originally conceived of through the lens of HOSVD (Tucker, 1966; De Lathauwer *et al.*, 2000). Specifically, structuring the parameters for different predictor combinations as a $x_{1,\max} \times \cdots \times x_{p,\max}$ dimensional tensor $\boldsymbol{\beta}_k = \{\beta_{k,x_1,\ldots,x_p} : (x_1,\ldots,x_p) \in \mathcal{X}\}$ for different predictor combinations at each location k and then applying an HOSVD-type factorization (Figure 4.4), model (4.3) can be reformulated as

$$\{\beta_{k,x_1,\ldots,x_p} \mid z_{j,k}^{(x_j)}, j = 1,\ldots,p\} = \textstyle\sum_{z_{1,k}} \cdots \sum_{z_{p,k}} \beta^{\star}_{k,z_{1,k},\ldots,z_{p,k}} \prod_{j=1}^{p} \mathbb{1}\{z_{j,k}^{(x_j)} = z_{j,k}\},$$

where $\boldsymbol{\beta}^{\star}_k = \{\beta^{\star}_{k,z_{1,k},\ldots,z_{p,k}} : (z_{1,k},\ldots,z_{p,k}) \in \mathcal{Z}_k\}$ is a $\ell_{1,k} \times \cdots \times \ell_{p,k}$ dimensional core tensor and $\mathbf{z}_{j,k} = \{\mathbb{1}\{z_{j,k}^{(x_j)} = z_{j,k}\} : x_j \in \mathcal{X}_j, z_{j,k} \in \mathcal{Z}_{j,k}\}$ are $x_{j,\max} \times \ell_{j,k}$ dimensional allocation matrices. In doing so, the problem of modeling the original parameter tensors $\boldsymbol{\beta}_k$ is effectively reduced to that of modeling the smaller-dimensional core tensors $\boldsymbol{\beta}^{\star}_k$. Significant reduction in model size is achieved at the location k when $\prod_{j=1}^{p} \ell_{j,k} \ll \prod_{j=1}^{p} x_{j,\max}$. Dynamic time-varying dependency structures appropriate for longitudinal settings are then accommodated via fHMM priors on the allocation variables $\mathbf{z}_{j,k}$ and novel Markovian priors on the core tensors $\boldsymbol{\beta}^{\star}_k$. The varying sizes and structures of the core tensors $\boldsymbol{\beta}^{\star}_k$ at different locations k (Figure 4.4) crucially allow the model to identify different sets of important predictors at different locations k. The second layer clustering of the elements of the core tensors

$$\beta^{\star}_{k,z_{1,k},\ldots,z_{p,k}} = \textstyle\sum_{z_k=1}^{\ell_k} \beta^{\star\star}_{k,z_k} \mathbb{1}\{z_k^{(z_{1,k},\ldots,z_{p,k})} = z_k\}$$

further refines the model (Figure 4.1), making the final partition structure fully flexible.

The HOSVD characterizes each x_j's overall significance explicitly and their joint influences implicitly but very compactly, efficiently eliminating the redundant variables and achieving significant reduction in dimensions, but avoids explicitly describing their main and lower-dimensional interaction effects (Johndrow *et al.*, 2017) which are often very useful to practitioners for their easy interpretation. These effects may, however, be meaningfully *defined* (and easily estimated from the posterior samples) directly as

$$\begin{aligned}
&\text{overall mean: } f_0(t) = \frac{\sum_{\mathbf{x}} f_{x_1,\ldots,x_p}(t)}{|\mathcal{X}|},\\
&\text{main effects: } f_{x_j}(t) = \frac{\sum_{\mathbf{x}_{-j}} f_{x_1,\ldots,x_p}(t)}{|\mathcal{X}_{-j}|} - f_0(t),\\
&\text{interactions: } f_{x_{j_1},x_{j_2}}(t) = \frac{\sum_{\mathbf{x}_{-j_1,-j_2}} f_{x_1,\ldots,x_p}(t)}{|\mathcal{X}_{-j_1,-j_2}|} - f_{x_{j_1}}(t) - f_{x_{j_2}}(t) - f_0(t), \text{ etc.},
\end{aligned} \tag{4.5}$$

where $\mathbf{x}_{-j} = (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p)^{\mathrm{T}} \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_{j-1} \times \mathcal{X}_{j+1} \times \cdots \times \mathcal{X}_p = \mathcal{X}_{-j}$, and so on.

Our proposed HOSVD based methodology provides a straightforward way to test for the presence of local overall effects of different predictors x_j using their marginal posterior inclusion probabilities (Figure 4.7), with consistency proven in Theorem 3.

When a predictor is found important overall, interest may additionally lie in testing the significance of its main and lower order interaction effects. We do not pursue the problem in more detail in this work but provide a general recipe for doing this by testing their pairwise differences here. For the main effects of predictor

$x_j \in \mathcal{X}_j = \{1, \ldots, x_{j,\max}\}$, for example, we may consider, at each time point t, $\binom{x_{j,\max}}{2}$ pairwise tests of the type

$$H_{0,j,\ell_1,\ell_2}(t) : \left| f_{x_{j\ell_1}}(t) - f_{x_{j\ell_2}}(t) \right| \leq \Delta_j(t) \quad \text{vs.} \quad H_{1,j,\ell_1,\ell_2}(t) : \left| f_{x_{j\ell_1}}(t) - f_{x_{j\ell_2}}(t) \right| > \Delta_j(t),$$

where we have followed (Berger, 1985, Chapter 4, pp. 148) in replacing conventional point nulls by reasonable interval nulls. If $H_{0,j,\ell_1,\ell_2}(t)$ is rejected in favor of $H_{1,j,\ell_1,\ell_2}(t)$ for some $\ell_1 \neq \ell_2$, we can conclude that the main effects of x_j are significant at time t. The interaction effects can be similarly tested.

Letting $d_{j,\ell_1,\ell_2}(t) = 1\left[\Pi\{H_{1,j,\ell_1,\ell_2}(t)|\text{data}\} > \beta\right]$ denote the decision rule, the posterior FDR (Chandra and Bhattacharya, 2019), defined as

$$\text{FDR}_{post} = \frac{\sum d_{j,\ell_1,\ell_2}(t)\Pi\{H_{0,j,\ell_1,\ell_2}(t)|\text{data}\}}{\max\{\sum d_{j,\ell_1,\ell_2}(t), 1\}},$$

can then be controlled at the level $1-\beta$ (Müller *et al.*, 2004). For a fixed β, the FDR_{post} depends on the choice of Δ. To obtain the optimal Δ, we can compute posterior FDR_{post}'s on a grid of Δ values in $(0,1)$ and set $\Delta = \inf_{\Delta'} \text{FDR}_{post}(\Delta') \leq 1-\beta$.

The HOSVD view is relevant particularly for the extremely challenging multivariate predictor problem $(x_1, \ldots, x_p)$ but not for a single predictor x, in which case the fHMM (Figure 4.2, right panel) simplifies to an HMM with a single layer $z_k^{(x)}$, and the second layer clustering of the spline coefficients $\beta^{\star}_{k,z_k^{(x)}}$ (Figure 4.1) is not needed. As discussed in the Introduction, the focus of the chapter is primarily on the multivariate case. Our implementation, however, is automated to adjust to both scenarios.

4.2.2 Random Effects Model

We model the random effects components $u_i(t)$ as

$$\begin{aligned} &u_i(t) = \textstyle\sum_{k=1}^{K} \beta_{k,i}^{(u)} B_k(t), \\ &\boldsymbol{\beta}_i^{(u)} \sim \mathrm{MVN}_K\{\mathbf{0}, (\sigma_{u,a}^{-2}\mathbf{I}_K + \sigma_{u,s}^{-2}\mathbf{P}_u)^{-1}\}, \\ &\sigma_{u,s} \sim \mathrm{C}^+(0, s_\sigma), \qquad \sigma_{u,a} \sim \mathrm{C}^+(0, s_\sigma), \end{aligned} \tag{4.6}$$

where $\boldsymbol{\beta}_i^{(u)} = (\beta_{1,i}^{(u)}, \ldots, \beta_{K,i}^{(u)})^{\mathrm{T}}$ are subject specific spline coefficients, $\mathrm{MVN}_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a K dimensional multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. We choose $\mathbf{P}_u = \mathbf{D}_u^{\mathrm{T}}\mathbf{D}_u$, where the $(K-1) \times K$ matrix $\mathbf{D}_u$ is such that $\mathbf{D}_u\boldsymbol{\beta}_i^{(u)}$ computes the first order differences in $\boldsymbol{\beta}_i^{(u)}$. The model thus penalizes $\sum_{k=1}^{K}(\nabla\beta_{k,i}^{(u)})^2 = \boldsymbol{\beta}_i^{(u)\mathrm{T}}\mathbf{P}_u\boldsymbol{\beta}_i^{(u)}$, the sum of squares of first order differences in $\boldsymbol{\beta}_i^{(u)}$ (Eilers and Marx, 1996). This induces a first order Markov dynamics for the spline coefficients, evident from the tridiagonal structure of the precision matrix in (4.6) that encodes their conditional dependence relationships. The random effects variance parameter $\sigma_{u,s}^2$ models the smoothness of the random effects curves, smaller $\sigma_{u,s}^2$ inducing smoother $u_i(t)$'s. Additional variations from the constant zero curve are explained by $\sigma_{u,a}^2$ (Figure 4.5). The absence of random effects is signified by the limiting case $\sigma_{u,s}^2 = \sigma_{u,a}^2 = 0$.

A similar model for functional random effects but with additional assumptions on the covariance matrix has previously been developed in Guo (2002). If we ignore the sharing of information through model hierarchies, the data for estimating an individual level effect come from that individual alone whereas the data for estimating the fixed effects come from many individuals with shared predictor levels. In the

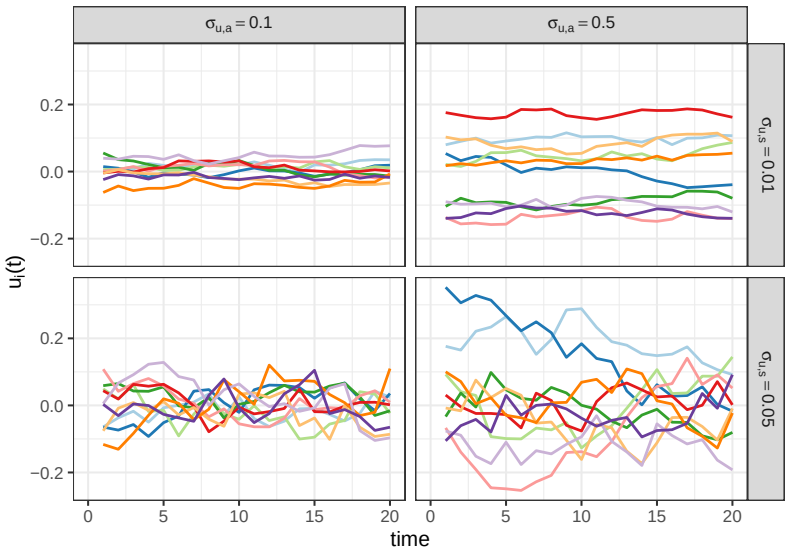


Figure 4.5: An illustration of the functional random effects model proposed in Section 4.2.2. Each panel shows a collection of 10 random draws from the random effects distribution for a combination of values of $(\sigma_{u,s}, \sigma_{u,a})$.

literature on mixed models, the random effects are thus often kept much simpler compared to the associated fixed effects models. In similar vein, we have focused here on time-varying random intercept type models. When categorical covariates, say $x'_1, \ldots, x'_{p'}$, are desired to be included in the random effects model, $u_i(t)$ can be modified as $u_{x'_1,\ldots,x'_{p'},i}(t)$ and the modeling strategies for the fixed effects components described in Section 4.2.1 can potentially be used.

4.3 Posterior Consistency

This section presents some convergence results for our proposed longitudinal functional mixed model. We focus on the case where $n \to \infty$ but L, the number

of replicates per individual, and T, the number of data recording time points, are kept fixed, which constitutes an appropriate asymptotic regime for the applications discussed later. Under this framework, we focus mainly on the recovery of the fixed effects components. When $L \to \infty$, similar results can also be established for the individual specific effects. We restrict ourselves to consistency at the knot points which coincide with the set of unique data observing time points in the setting of this work. The functional domain remaining fixed to a finite interval, say $[A, B]$, when the number of data recording time points inside the domain $T \to \infty$ and some additional mild smoothness assumptions are made on the true underlying functions, the results can also be extended to the entire domain.

Our proofs rely on some results and ideas from Ghosal *et al.* (1999) and Suarez and Ghosal (2016). We first show consistency for the functional fixed effects. Using this result, we then show that our proposed model can also recover the underlying true local partitions of the covariate space and hence perform consistent variable selection.

We let $\Pi(\cdot)$ denote the prior distribution induced by our model on the space of fixed effects functions $f_{\mathbf{x}}(t)$ and $\Pi(\cdot|\text{data})$ denote the corresponding posterior. We let $g(\mathbf{x})$ denote the probability distribution of $\mathbf{x}$. We consider the g-weighted local L_2-norm of the function $f_{\mathbf{x}}(t)$, defined as $||f||^2_{2,g,loc} = \sum_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}) \sum_{k=1}^{K} f^2_{\mathbf{x}}(k)$. For the linear B-spline mixtures used in this work, $f_{\mathbf{x}}(k) = \beta_{k,\mathbf{x}}$.

Integrating out the random effects distribution (4.6) from model (4.1), we obtain

$$\{y_{i,\ell,t} \mid x_{j,i,\ell,t} = x_j, j = 1, \ldots, p\} \sim \text{Normal}\{f_{\mathbf{x}}(t), \sigma^2_\varepsilon + \sigma^2_u(t)\}, \tag{4.7}$$

where $\sigma_u^2(t) = \{(\sigma_{u,a}^{-2}\mathbf{I}_K + \sigma_{u,s}^{-2}\mathbf{P}_u)^{-1}\}_{t,t}$. In our proof, we deviate slightly from our stated model in assuming exponentially decaying tails for the priors on the variance parameters $\sigma_{u,a}^2$ and $\sigma_{u,s}^2$ instead of the more non-informative half-Cauchy priors we used in our implementation.

Theorem 1 (function estimation)**.** *For any* $\epsilon > 0$, $\Pi(||f - f_0||_{2,g,loc} < \epsilon \mid \text{data}) \to 1$.

Proof. *For notational convenience, we let* $f_{k,\mathbf{x},0} = f_{\mathbf{x},0}(k)$ *be denoted, without loss of generality, by* $\beta_{k,\mathbf{x},0}$*. As our focus is on the parameters* f *of a conditional probability distribution of the type* $p(y|f,\mathbf{x})$*, we can fix the marginal distribution* $g(\mathbf{x})$ *of* $\mathbf{x}$ *at its true value, say* $g_0(\mathbf{x})$*, and model the unknown conditional distribution* $p(y|f,\mathbf{x})$ *independently of* $g(\mathbf{x})$*. We can simply restrict ourselves to the set of joint distributions such that* $p(y,\mathbf{x}|f) = p(y|f,\mathbf{x})g_0(\mathbf{x})$*. With some abuse of notation, we have thus not distinguished between* $g(\mathbf{x})$ *and* $g_0(\mathbf{x})$*, nor between the joint distribution of* $(y,\mathbf{x}|f)$ *and the conditional distribution of* $(y|f,\mathbf{x})$ *but denote them both by* p_f*.*

We start by proving that the true data generating density p_{f_0} *is in the Kullback-Leibler support of the prior* Π*, i.e.,* $\Pi\left(d_{KL}(p_f, p_{f_0}) < \epsilon\right) > 0\ \forall \epsilon > 0$*. The KL support property shows the theoretical flexibility of the proposed model in encompassing a large class of true data generating scenarios. We note that*

$$\begin{aligned}
&\Pi\left(\|f - f_0\|_{2,g,loc}^2 < \epsilon^2\right) = \Pi\left(\textstyle\sum_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}) \sum_{k=1}^K |\beta_{k,\mathbf{x}} - \beta_{k,\mathbf{x},0}|^2 < \epsilon^2\right) \\
&\geq \Pi\left\{|\beta_{k,\mathbf{x}} - \beta_{k,\mathbf{x},0}|^2 < \epsilon^2/(K\,|\mathcal{X}|\,\delta_{g,\max}),\ \forall \mathbf{x}, \forall k\right\} \\
&\geq \Pi\left\{|\beta_{k,z}^{\star\star} - \beta_{k,z,0}^{\star\star}|^2 < \epsilon^2/(K\,|\mathcal{X}|\,\delta_{g,\max}),\ \forall \mathbf{x}, \forall k \mid A\right\} \Pi(A),
\end{aligned}$$

where $\delta_{g,\max} = \max_{\mathbf{x}} g(\mathbf{x})$*, the event* $A = \{z_{j,k}^{(x_j)} = x_j\ \forall j,k; z_k^{(x_1,\ldots,x_p)} = z;\ m_k = \ell_k\}$ *denotes a special case when no clustering occurs at any time point, or, in other words,*

the different possible level combinations of $\mathbf{x}$ *all form their own separate clusters at all time points. By the construction of our partition model, the event* A *has a positive prior probability. It is possible to explicitly calculate the prior on the spline coefficients, conditional on the event* A*. Specifically, we get*

$$\begin{pmatrix} \beta^{\star\star}_{1,\mathbf{x}} \\ \beta^{\star\star}_{2,\mathbf{x}} \\ \vdots \\ \beta^{\star\star}_{K-1,\mathbf{x}} \\ \beta^{\star\star}_{K,\mathbf{x}} \end{pmatrix} \sim MVN_K \left\{ \begin{pmatrix} \mu_0 \\ \mu_0 \\ \vdots \\ \mu_0 \\ \mu_0 \end{pmatrix}; \begin{pmatrix} \sigma_\beta^{-2}+\sigma_0^{-2} & -\sigma_\beta^{-2} & & & \\ -\sigma_\beta^{-2} & 2\sigma_\beta^{-2} & -\sigma_\beta^{-2} & & \\ & -\sigma_\beta^{-2} & \ddots & \ddots & \\ & & \ddots & \ddots & -\sigma_\beta^{-2} \\ & & & -\sigma_\beta^{-2} & \sigma_\beta^{-2} \end{pmatrix}^{-1} \right\}.$$

The above result follows by assuming that $\beta^{\star\star}_{1,\mathbf{x}} \sim Normal(\mu_0, \sigma_0^2)$*, which approximates the setting of this work in the limiting case* $\sigma_0^2 \to \infty$*. When* σ_0^2 *is finite, the precision matrix of the prior on the spline coefficients is symmetric positive definite. Since the joint distribution of the spline coefficients has full support on* $\mathbb{R}^K$*, it follows that* $\Pi\left(\|f - f_0\|_{2,g,loc} < \epsilon\right) > 0$*. This shows the positivity of any Kullback-Leibler neighborhood since in the case of a Gaussian likelihood* $d_{KL}(p_f, p_{f_0}) = \|f - f_0\|^2_{2,g,loc}/(2\sigma_n^2)$*.*

To establish strong consistency for the posterior distribution of $f(t)$*, we apply Theorem 2 of (Ghosal* et al.*, 1999) stated below for easy reference.*

Theorem 2 (Ghosal et al., 1999)**.** *Let* Π *be a prior on* $\mathcal{F}$*. Suppose* $p_{f_0} \in \mathcal{F}$ *is in the KL support of* Π *and let* $U = \{p_f : \|p_f - p_{f_0}\| < \epsilon\}$*, where* $\|\cdot\|$ *is the* L_1*-norm for the densities. If there is a* $\delta < \epsilon/4, c_1, c_2 > 0, \alpha < \epsilon^2/8$ *and* $\mathcal{F}_n \subset \mathcal{F}$ *such that, for all* n *large:*

(i) $\Pi(\mathcal{F}_n^c) < c_1 \exp(-nc_2)$, *and*

(ii) $J(\delta, \mathcal{F}_n, \|\cdot\|) < n\alpha$,

then $\Pi(U \mid data) \to 1$.

For $\delta > 0$, *the metric entropy* $J(\delta, \mathcal{F}_n, \|\cdot\|)$ *is the logarithm of the minimum of all* k *such that there exist* $p_1, p_2, \ldots, p_k \in \mathcal{F}$ *with the property* $\mathcal{F}_n \subset \cup_{i=1}^k \{p : \|p - p_i\| < \delta\}$. *We can construct a sieve in the parameter space* $\mathcal{H}_n = \{\boldsymbol{\theta} : ||\boldsymbol{\beta}^{\star\star}||_\infty < M_{1n}, m_{2n} < \sigma_n^2 < M_{2n}\}$ *and* $\mathcal{F}_n = \{p_f : \boldsymbol{\theta} \in \mathcal{H}_n\}$.

We have already verified the KL support condition. Similarly, we can use results for the Gaussian likelihood to bound the L_1 *distance as* $\|p_f - p_{f_0}\| \leq C_1 \frac{\|f - f_0\|_\infty}{\sigma_n}$ *for some constant* C_1. *The logarithm of the minimum number of brackets of size* δ *required to cover* $\mathcal{F}_n$ *is bounded as*

$$\begin{aligned} J(\delta, \mathcal{F}_n, \|\cdot\|) &< J(m_{2n}\delta/C_1, \{\boldsymbol{\beta}^{\star\star}, \sigma_n^2 : \|\boldsymbol{\beta}^{\star\star}\|_\infty < M_{1n}, m_{2n} < \sigma_n^2 < M_{2n}\}, \|\cdot\|_\infty) \\ &< K \log\{3C_1 K M_{1n}/(\delta m_{2n})\}. \end{aligned}$$

We need to analyze the tail behavior of $\sigma_n^2(t) = \sigma_\varepsilon^2 + \sigma_u^2(t)$ *to find a bound for the sieve complement. Using Cramer's rule to calculate the inverse matrix, it is easy to see that* $\sigma_u^2(t) = \{(\sigma_{u,a}^{-2}\mathbf{I}_K + \sigma_{u,s}^{-2}\mathbf{P}_u)^{-1}\}_{t,t} = O(\sigma_{u,a}^2)$. *Thus,* $\sigma_n^2 = O(\sigma_\varepsilon^2) + O(\sigma_{u,a}^2)$. *The prior probability of the sieve-complement can then be bounded as*

$$\begin{aligned} \Pi(\mathcal{H}_n^c) &< \Pi(\boldsymbol{\beta}^{\star\star} \notin [-M_{1n}, M_{1n}]^K) + \Pi(\sigma_n^2 \notin [m_{2n}, M_{2n}]) \\ &< K \exp(-R_1 M_{1n}^{t_1}) + \exp(-R_2 M_{2n}^{t_2}) \end{aligned} \tag{4.8}$$

for some constants R_1, t_1, R_2, t_2 *and some sequences* M_{1n}, m_{2n}, M_{2n}*. In order to apply Theorem 2, we then need, for* $\delta < \epsilon/4$ *and* $\alpha < \epsilon^2/8$*, that*

$$\begin{aligned} & K \ \log\{3C_1 K M_{1n}/(\delta m_{2n})\} < n\alpha \quad \textit{and} \\ & K \exp(-R_1 M_{1n}^{t_1}) + \exp(-R_2 M_{2n}^{t_2}) < c_1 \exp(-nc_2). \end{aligned} \tag{4.9}$$

Conditions 4.8 and 4.9 are satisfied by choosing M_{1n} *to be a positive polynomial of* n*,* m_{2n} *be a negative polynomial of* n*, and* M_{2n} *be an exponential function of* n*, depending appropriately on the constants in these equations. Hence, we have* $\Pi(||p_f - p_{f_0}|| < \epsilon \mid data) \to 1$*. We conclude that the posterior distribution is consistent relative to the* L_1 *distance. Finally, since* $\|f - f_0\|_{2,g,loc} \lesssim \|p_f - p_{f_0}\|$ *we get* $\Pi(\|f - f_0\|_{2,g,loc} < \epsilon \mid data) \to 1$. ■

Without any loss of generality, we assume that $g(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{X}$. If not, we can simply restrict ourselves to the set on which $g(\mathbf{x}) > 0$. We then have $n_{\mathbf{x}} \to \infty$ as $n \to \infty$ for all $\mathbf{x} \in \mathcal{X}$. The asymptotic regime can then be understood as averaging over $n_{\mathbf{x}}$ replications for each $\mathbf{x}$, thus replacing $\sigma_\varepsilon^2 + \sigma_u^2(t)$ by $\sigma_{n,\mathbf{x}}^2 = n_{\mathbf{x}}^{-1}\{\sigma_\varepsilon^2 + \sigma_u^2(t)\}$. Theorem 1 then implies that, for any $\mathbf{x} \in \mathcal{X}$ and $\epsilon > 0$, $\Pi(||f_{\mathbf{x}} - f_{\mathbf{x},0}||_{2,loc} < \epsilon \mid \text{data}) \to 1$, where $\mathbf{f}_{\mathbf{x}} = (f_{1,\mathbf{x}}, \ldots, f_{K,\mathbf{x}})^{\mathrm{T}}$ with $f_{k,\mathbf{x}} = f_{\mathbf{x}}(k)$ and $||f_{\mathbf{x}}||_{2,loc}^2 = \sum_{k=1}^K f_{\mathbf{x}}^2(k)$.

For a given location k, let $\boldsymbol{\rho}_k = \{\boldsymbol{\rho}_{k,1}, \ldots, \boldsymbol{\rho}_{k,m_k}\}$ be a random partition of $\mathcal{X}$, the space of vectors of length p whose individual entries have values in $\mathcal{X}_j$, respectively. The partition $\boldsymbol{\rho}_k$ is defined in the following way:

$$\beta_{k,\mathbf{x}}^{\star\star} = \beta_{k,\mathbf{x}'}^{\star\star} \Longleftrightarrow \mathbf{x}, \mathbf{x}' \in \boldsymbol{\rho}_{k,h} \text{ for some } h \in \{1, \ldots, m_k\}.$$

Our hierarchical prior for the random partitions assigns a positive probability to each possible configuration. Let $\boldsymbol{\rho}_{k,0}$ be the partition generated by the true values of the parameters at location k. Then the following theorem holds.

Theorem 3 (variable selection)**.** $\Pi(\boldsymbol{\rho}_k = \boldsymbol{\rho}_{k,0} \mid \text{data}) \to 1$.

Proof. *Let a compatible model be a collection of all parameter values corresponding to a partition which is finer than $\boldsymbol{\rho}_{k,0}$. Let an incompatible model be any collection of parameters that does not result in a compatible model.*

Theorem 1 implies that the posterior probability for any neighborhood of the true value $\beta_{k,\mathbf{x},0}$ for the spline coefficients converges to 1*. For incompatible models, there exists an incorrect assignment for a pair $(\mathbf{x}, \mathbf{x}')$, i.e., $\beta_{k,\mathbf{x}} = \beta_{k,\mathbf{x}'}$ when $\beta_{k,\mathbf{x},0} \neq \beta_{k,\mathbf{x}',0}$. Therefore, is it possible to find a neighborhood of $\beta_{k,\mathbf{x},0}$ that does not contain $\beta_{k,\mathbf{x}}$, which contradicts Theorem 1.*

Hence, we can focus exclusively on the set of compatible models. Let $\boldsymbol{\rho}_k = \{\boldsymbol{\rho}_{k,1}, \ldots, \boldsymbol{\rho}_{k,m_k}\}$ be a compatible model and $\boldsymbol{\rho}_{k,0} = \{\boldsymbol{\rho}_{k,1,0}, \ldots, \boldsymbol{\rho}_{k,m_{k,0},0}\}$ be the true model. Since $\boldsymbol{\rho}_k$ is a finer partition, we can assume without loss of generality that $m_k > m_{k,0}$ and that $\boldsymbol{\rho}_{k,h} \subseteq \boldsymbol{\rho}_{k,h,0}$ for $h = 1, \ldots, m_{k,0}$. In Section 4.4, we calculated the marginal likelihood of the model as

$$\begin{aligned} p(\mathbf{y}_k \mid \boldsymbol{\rho}_k, \boldsymbol{\zeta}) &= \prod_{h=1}^{m_k} (2\pi\sigma_n^2)^{-\frac{n_{k,h}}{2}} (\sigma_{k,h}^2)^{-\frac{1}{2}} (\sigma_{k,h}^{\star 2})^{\frac{1}{2}} e^{-\frac{1}{2}\left(\frac{\sum_{i,\ell} y_{i,\ell,k}^2}{\sigma_n^2} + \frac{\mu_{k,h}^2}{\sigma_{k,h}^2} - \frac{\mu_{k,h}^{\star 2}}{\sigma_{k,h}^{\star 2}}\right)} \\ &= \prod_{h=1}^{m_k} (2\pi\sigma_n^2)^{-\frac{n_{k,h}}{2}} \left(1 + n_{k,h}\sigma_n^{-2}\sigma_{k,h}^2\right)^{-\frac{1}{2}} e^{-\frac{1}{2}\left(\frac{\sum_{i,\ell} y_{i,\ell,k}^2}{\sigma_n^2} + \frac{\mu_{k,h}^2}{\sigma_{k,h}^2} - \frac{\mu_{k,h}^{\star 2}}{\sigma_{k,h}^{\star 2}}\right)}. \end{aligned}$$

Now,

$$\frac{p(\mathbf{y}_k \mid \boldsymbol{\rho}_k, \boldsymbol{\zeta})}{p(\mathbf{y}_k \mid \boldsymbol{\rho}_{k,0}, \boldsymbol{\zeta})} = \sqrt{\frac{\prod_{h=1}^{m_{k,0}}(1 + n_{k,h,0}\sigma_n^{-2}\sigma_{k,h,0}^2)}{\prod_{h=1}^{m_k}(1 + n_{k,h}\sigma_n^{-2}\sigma_{k,h}^2)}} \exp\left\{-\frac{1}{2}\sum_{h=1}^{m_k}\left(\frac{\mu_{k,h}^2}{\sigma_{k,h}^2} - \frac{\mu_{k,h}^{\star 2}}{\sigma_{k,h}^{\star 2}}\right) + \frac{1}{2}\sum_{h=1}^{m_{k,0}}\left(\frac{\mu_{k,h,0}^2}{\sigma_{k,h,0}^2} - \frac{\mu_{k,h,0}^{\star 2}}{\sigma_{k,h,0}^{\star 2}}\right)\right\}.$$

To examine the behavior of this expression, note that $m_k > m_{k,0}$, *and* $n_{k,h} < n_{k,h,0}$, $h = 1, \ldots, m_{k,0}$. *Now, the expression under the square root converges to* 0 *as* $n \to \infty$ *since* $m_k > m_{k,0}$ *implies that the denominator is of a higher order.*

Let us now focus on the exponential term. Notice that both $\mu_{k,h}^2/\sigma_{k,h}^2$ *and* $\mu_{k,h,0}^2/\sigma_{k,h,0}^2$ *are* $O_n(1)$. *Moreover, the only terms in the exponential that are not* $O_n(1)$ *can be rewritten as*

$$-\frac{1}{2\sigma_n^2}\left[\sum_{h=1}^{m_k}\frac{\left(\sum_{\substack{(i,\ell) \text{ s.t.} \\ \mathbf{x}_{i,\ell,t} \in \boldsymbol{\rho}_{k,h}}} y_{i,\ell,k}\right)^2}{n_{k,h}} - \sum_{h=1}^{m_{k,0}}\frac{\left(\sum_{\substack{(i,\ell) \text{ s.t.} \\ \mathbf{x}_{i,\ell,t} \in \boldsymbol{\rho}_{k,h,0}}} y_{i,\ell,k}\right)^2}{n_{k,h,0}}\right].$$

The expression in the square brackets can be rewritten as a quadratic form $\mathbf{y}^\intercal\mathbf{A}\mathbf{y}$ *for some matrix* $\mathbf{A}$ *which is* $O_n(1)$. *Since the dispersion matrix of* $\mathbf{y}$ *is* $\sigma_n^2\mathbf{I}_n$, *the variance of* $\mathbf{y}^\intercal\mathbf{A}\mathbf{y}$ *is* $O_n(1/n)$ *as* $n \to \infty$. *Therefore, the exponential term is bounded in probability as* $n \to \infty$.

Thus, we have shown that the probability of any incompatible model goes to zero. Along with the fact that for any compatible model the marginal likelihood ratio tends to 0 *implies that the only model that can possibly retain positive probability is the truth. Since for any fixed* k *there are only finitely many models for* $\{\beta_{k,\mathbf{x}}\}_{\mathbf{x}\in\mathbb{X}}$, *the probability of the true model must tend to* 1. ■

The construction of our model in Section 4.2.1 is such that the influences of the predictors are encoded precisely by the model induced partition structures - the predictor x_j is important at location k if its levels belong to at least two different sets in the partition $\boldsymbol{\rho}_k$. Consistency in recovering the local partitions thus immediately implies consistency in local variable selection.

We have not studied posterior contraction rates here which would require quantifying bounds for the prior concentration around the truth, the entropy of the model, etc. (Ghosal and Van der Vaart, 2017) under the complex dependence relationships induced by our model via the many sophisticated layers of hierarchy discussed in Section 4.2, daunting tasks that we view to be outside the scope of this work.

4.4 Posterior Inference

Inference for the proposed LFMM is based on samples drawn from the posterior using an MCMC algorithm. In our model, the values of $\ell_{j,k}$'s are crucial in controlling the model size since they act as local covariate importance indicators. Varying values of $\ell_{j,k}$'s, however, result in varying dimensional models, posing daunting computational challenges. Dynamic message passing algorithms, such as the forward-backward sampler, are popular strategies for inference in HMMs and fHMMs (Rabiner, 1989; Scott, 2002). However, it is not clear how message passing strategies can be adapted to include inferences about the $\ell_{j,k}$'s.

We address these challenges by designing an efficient trans-dimensional transition step which updates the partition structure and the corresponding local curves

at every location. First, for every location k, an update in the partition structure $\boldsymbol{\rho}_k$ is proposed. Second, conditional on $\boldsymbol{\rho}_k$, samples of the spline coefficients $\boldsymbol{\beta}_k^{\star\star} = \{\beta_{k,h}^{\star\star}\}_{h=1}^{m_k}$ are drawn from their Gaussian full conditional distributions.

Specifically, the first step involves updating, for every predictor j at each location k, the first layer of latent variables $\mathbf{z}_{j,k} = (z_{j,k}^{(1)}, \ldots, z_{j,k}^{(x_{j,\max})})$, the implied partition sizes $(\ell_{1,k}, \ldots, \ell_{p,k})$, and the corresponding second layer of latent variables $\mathbf{z}_k = z_k^{(z_{1,k},\ldots,z_{p,k})}$. Designing an efficient such proposal is made challenging by the discrete and potentially high-dimensional support of the latent variables $\mathbf{z}_{j,k}$ and $\mathbf{z}_k$. However, the proposal distribution can be defined sequentially as

$$q(\ell'_{j,k}, \mathbf{z}'_{j,k}, \mathbf{z}'_k \mid \ell_{j,k}, \mathbf{z}_{j,k}, \mathbf{z}_k) = q_1(\ell'_{j,k}, \mathbf{z}'_{j,k} \mid \ell_{j,k}, \mathbf{z}_{j,k}) q_2(\mathbf{z}'_k \mid \ell'_{j,k}, \mathbf{z}'_{j,k}).$$

First, we perturb the current state $\mathbf{z}_{j,k}$ to a new configuration $\mathbf{z}'_{j,k}$ by sampling uniformly in a Hamming ball of radius r around $\mathbf{z}_{j,k}$ (Titsias and Yau, 2014), resulting in an efficient first layer proposal that shares many of the old components as

$$q_1(\ell'_{j,k}, \mathbf{z}'_{j,k} \mid \ell_{j,k}, \mathbf{z}_{j,k}) = \text{Unif}\{\mathbf{z}'_{j,k} \mid \mathcal{H}_m(\mathbf{z}_{j,k})\} \mathbb{1}\{\ell'_{j,k} = |\mathcal{Z}'_{j,k}|\}.$$

Conditioning on the first layer of latent variables, we update the second layer as

$$q_2(\mathbf{z}'_k \mid \ell'_{j,k}, \mathbf{z}'_{j,k}) = \text{Mult}\left(1/\ell_k, \ldots, 1/\ell_k\right).$$

In terms of the implied marginal partition structure $\mathcal{Z}_{j,k}$, when $r = 1$, this corresponds to (A) selecting a covariate level and either (Ba) merging it to one of the other existing partition elements or (Bb) creating a singleton by separating it from its partition element. Since the first layer proposal distribution is symmetric, the

resulting acceptance rate of the Metropolis-Hastings (M-H) step is

$$r_{acc} = \frac{p(\mathbf{y}_k \mid \boldsymbol{\rho}'_k, \sigma^2_\varepsilon, \sigma^2_\beta, \boldsymbol{\zeta})}{p(\mathbf{y}_k \mid \boldsymbol{\rho}_k, \sigma^2_\varepsilon, \sigma^2_\beta, \boldsymbol{\zeta})} \cdot \frac{p(\mathbf{z}'_{j,k})p(\ell'_{j,k})p(\mathbf{z}'_k)}{p(\mathbf{z}_{j,k})p(\ell_{j,k})p(\mathbf{z}_k)} \cdot \frac{q_2(\mathbf{z}_k \mid \ell'_{j,k}, \mathbf{z}'_{j,k})}{q_2(\mathbf{z}'_k \mid \ell'_{j,k}, \mathbf{z}'_{j,k})}, \tag{4.10}$$

where $\boldsymbol{\zeta}$ denotes a generic variable that collects all other variables not explicitly mentioned here, including the data points, and $\mathbf{y}_k = \{y_{i,\ell,k}\}_{i,\ell}$. Importantly, the spline coefficient parameters $\boldsymbol{\beta}^{\star\star}_k$ at each location k can be analytically integrated out of the posterior of the corresponding partition structure. This allows for an efficient scheme for sampling the random partition structures based on their marginal likelihood

$$p(\mathbf{y}_k \mid \boldsymbol{\rho}_k, \sigma^2_\varepsilon, \sigma^2_\beta, \boldsymbol{\zeta}) = \prod_{h=1}^{m_k} \int p(\mathbf{y}_k \mid \beta^{\star\star}_{k,h}, \boldsymbol{\rho}_{k,h}, \sigma^2_\varepsilon) p(\beta^{\star\star}_{k,h} \mid \sigma^2_\beta, \boldsymbol{\zeta}) d\beta^{\star\star}_{k,h}.$$

The second term in the integral is the conditional smoothing prior for the spline coefficients

$$\begin{aligned} p(\beta^{\star\star}_{k,h} \mid \sigma^2_\beta, \boldsymbol{\zeta}) &\propto \prod_{h' \in \mathcal{Z}^-_{k,h}} \text{Normal}\left(\beta^{\star\star}_{k,h} \mid \beta^{\star\star}_{k-1,h'}, \sigma^2_\beta\right) \prod_{h'' \in \mathcal{Z}^+_{k,h}} \text{Normal}\left(\beta^{\star\star}_{k+1,h''} \mid \beta^{\star\star}_{k,h}, \sigma^2_\beta\right) \\ &= \text{Normal}(\mu_{k,h}, \sigma^2_{k,h}), \end{aligned}$$

where $\mathcal{Z}^-_{k,h} = \left\{ z_{k-1} : z_{k-1} = z_{k-1}^{(z_{1,k-1},\ldots,z_{p,k-1})}; (z_{1,k-1}, \ldots, z_{p,k-1}) = (z_{1,k-1}^{(x_1)}, \ldots, z_{p,k-1}^{(x_p)}); (x_1, \ldots, x_p) \in \boldsymbol{\rho}_{k,h} \right\}$ and $\mathcal{Z}^+_{k,h} = \left\{ z_{k+1} : z_{k+1} = z_{k+1}^{(z_{1,k+1},\ldots,z_{p,k+1})}; (z_{1,k+1}, \ldots, z_{p,k+1}) = (z_{1,k+1}^{(x_1)}, \ldots, z_{p,k+1}^{(x_p)}); (x_1, \ldots, x_p) \in \boldsymbol{\rho}_{k,h} \right\}$ are the indexes of the coefficients expressed at the previous and following locations, respectively, $n^-_{k,h} = |\mathcal{Z}^-_{k,h}|$ and $n^+_{k,h} = |\mathcal{Z}^+_{k,h}|$ are the corresponding cardinalities, $\mu_{k,h} = \frac{\sum_{h'} \beta^{\star\star}_{k-1,h'} + \sum_{h''} \beta^{\star\star}_{k+1,h''}}{n^-_{k,h} + n^+_{k,h}}$ and $\sigma^2_{k,h} = \sigma^2_\beta (n^-_{k,h} + n^+_{k,h})^{-1}$ are the resulting smoothing prior mean and variance parameters. Hence we

get

$$
\begin{aligned}
&p(\mathbf{y}_k \mid \boldsymbol{\rho}_k, \sigma^2_\varepsilon, \sigma^2_\beta, \boldsymbol{\zeta}) \\
&= \prod_{h=1}^{m_k} \int \prod_{\substack{(i,\ell) \text{ s.t.} \\ \mathbf{x}_{i,\ell,t} \in \boldsymbol{\rho}_{k,h}}} \left\{ (2\pi\sigma^2_\varepsilon)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2_\varepsilon}(r^{(m)}_{i,\ell,k} - \beta^{\star\star}_{k,h})^2} \right\} (2\pi\sigma^2_{k,h})^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2_{k,h}}\left(\beta^{\star\star}_{k,h} - \mu_{k,h}\right)^2} d\beta^{\star\star}_{k,h} \\
&= \prod_{h=1}^{m_k} (2\pi\sigma^2_\varepsilon)^{-\frac{n_{k,h}}{2}} (\sigma^2_{k,h})^{-\frac{1}{2}} (\sigma^{\star 2}_{k,h})^{\frac{1}{2}} e^{-\frac{1}{2}\left(\frac{\sum_{i,\ell} r^{(m)2}_{i,\ell,k}}{\sigma^2_\varepsilon} + \frac{\mu^2_{k,h}}{\sigma^2_{k,h}} - \frac{\mu^{\star 2}_{k,h}}{\sigma^{\star 2}_{k,h}} \right)},
\end{aligned}
$$

where $\mathbf{r}^{(m)} = \{y_{i,\ell,t} - u_i(t)\}_{i,\ell,t}$ are the main effects residuals, $n_{k,h} = |\{(i,l) \text{ s.t. } \mathbf{x}_{i,\ell,t} \in \boldsymbol{\rho}_{k,h}\}|$ is the number of observations allocated to the spline coefficient $\beta^{\star\star}_{k,h}$, $\sigma^{\star 2}_{k,h} = \left(\sigma^{-2}_\varepsilon n_{k,h} + \sigma^{-2}_{k,h}\right)^{-1}$ and $\mu^\star_{k,h} = \sigma^{\star 2}_{k,h}\left(\sigma^{-2}_\varepsilon \sum_{i,\ell} r^{(m)}_{i,\ell,k} + \sigma^{-2}_{k,h}\mu_{k,h}\right)$.

Conditional on the partition structure $\boldsymbol{\rho}_k$, the group specific curves are sampled from their Gaussian full conditional distribution

$$
\begin{aligned}
p(\beta^{\star\star}_{k,h} \mid \mathbf{y}_k, \boldsymbol{\rho}_{k,h}, \sigma^2_\varepsilon, \sigma^2_\beta, \boldsymbol{\zeta}) &\propto p(\mathbf{y}_k \mid \beta^{\star\star}_{k,h}, \boldsymbol{\rho}_{k,h}, \sigma^2_\varepsilon) p(\beta^{\star\star}_{k,h} \mid \sigma^2_\beta, \boldsymbol{\zeta}) \\
&= \text{Normal}\left\{\mu^\star_{k,h}, \sigma^{\star 2}_{k,h}\right\}.
\end{aligned} \tag{4.11}
$$

To simplify posterior sampling for the scale parameter σ_β, we used a hierarchical scale mixture representation of the half-Cauchy distribution (Makalic and Schmidt, 2016). Introducing an auxiliary variable ν_β, the $\text{C}^+(0, s_\sigma)$ prior can be represented as

$$
\sigma^2_\beta \sim \text{Inv-Ga}(1/2, 1/\nu_\beta), \qquad \nu_\beta \sim \text{Inv-Ga}(1/2, 1/s^2_\sigma).
$$

Posterior full conditionals for the σ^2_β and ν_β then belong to the inverse-Gamma family and can be easily sampled from. The same trick, however, does not yield tractable full conditionals for $\sigma^2_{u,s}$ and $\sigma^2_{u,a}$. M-H steps are used for these parameters.

The full MCMC sampler comprises the steps reported in Algorithm 1 in Section 4.4.1. Our software implementation in `R` and `C++` is highly automated, requiring only the available data points and the values of a few prior hyper-parameters as inputs. These hyper-parameters appear deep inside the model hierarchy and inference is highly robust to their choices.

4.4.1 MCMC Algorithm

We summarize here the steps of the MCMC algorithm used to sample from the posterior of our model.

Algorithm 1

Updating the core tensor sizes $\ell_{j,k}$ and the latent variables $z_{j,k}^{(x_j)}, z_k^{(z_{1,k},\ldots,z_{p,k})}$

1: For $k = 1, \ldots, K$, $j = 1, \ldots, p$, sample $\ell_{j,k}$, $\mathbf{z}_{j,k}$ and $z_k^{(z_{1,k},\ldots,z_{p,k})}$ using the M-H step with acceptance rate (4.10).

Updating the cluster specific parameters $\beta_{k,h}^{\star\star}$ and the smoothness σ_β^2

2: For $k = 1, \ldots, K$, $h = 1, \ldots, m_k$ sample the group specific curves $\beta_{k,h}^{\star\star}$ from their Gaussian full conditionals (4.11)

$$p(\beta_{k,h}^{\star\star} \mid \mathbf{y}_k, \boldsymbol{\rho}_{k,h}, \sigma_\varepsilon^2, \sigma_\beta^2, \boldsymbol{\zeta}) = \text{Normal}\left\{\mu_{k,h}^{\star}, \sigma_{k,h}^{\star 2}\right\}.$$

3: Sample the smoothness parameter σ_β^2 from its inverse-Gamma full conditional

$$p(\sigma_\beta^2 \mid \boldsymbol{\zeta}) \propto p_0(\sigma_\beta^2) \prod_k \prod_h p(\beta_{k,h}^{\star\star} \mid \sigma_\beta^2).$$

Updating the initial distributions $\boldsymbol{\pi}_0^{(j)}$ and the transition dynamics $\boldsymbol{\pi}_h^{(j)}$

4: Let $n_{0,h}^{(j)} = \sum_{x_j} \mathbb{1}\{z_{j,1}^{(x_j)} = h\}$. For $j = 1, \ldots, p$, sample $\boldsymbol{\pi}_0^{(j)}$ as $\{\pi_0^{(j)}(1), \ldots, \pi_0^{(j)}(x_{j,\max})\} \mid \boldsymbol{\zeta} \sim \text{Dir}\{\alpha^{(j)}/x_{j,\max} + n_{0,1}^{(j)}, \ldots, \alpha^{(j)}/x_{j,\max} + n_{0,x_{j,\max}}^{(j)}\}$.

5: Let $n^{(j)}_{h,h'} = \sum_{k=2}^{K} \sum_{x_j} \mathbb{1}\{z^{(x_j)}_{j,k} = h, z^{(x_j)}_{j,k-1} = h'\}$. For $j = 1, \ldots, p$, $h = 1, \ldots, x_{j,\max}$, sample $\boldsymbol{\pi}^{(j)}_h$ as
$\{\pi^{(j)}_h(1), \ldots, \pi^{(j)}_h(x_{j,\max})\} \mid \boldsymbol{\zeta} \sim \text{Dir}\{\alpha^{(j)}/x_{j,\max} + n^{(j)}_{h,1}, \ldots, \alpha^{(j)}/x_{j,\max} + n^{(j)}_{h,x_{j,\max}}\}$.

6: For $j = 1, \ldots, p$, sample $\alpha^{(j)}$ using an M-H step from its full conditional
$$p(\alpha^{(j)} \mid \boldsymbol{\zeta}) \propto p_0(\alpha^{(j)}) p(\boldsymbol{\pi}^{(j)}_0 \mid \alpha^{(j)}) \prod_{h=1}^{x_{j,\max}} p(\boldsymbol{\pi}^{(j)}_h \mid \alpha^{(j)}).$$

Updating the second layer probabilities $\boldsymbol{\pi}^{\star}_k$

7: Let $n^{\star}_{k,h} = \sum_{z_{1,k},\ldots,z_{p,k}} \mathbb{1}\{z^{(z_{1,k},\ldots,z_{p,k})}_k = h\}$. Sample $\boldsymbol{\pi}^{\star}_k$ as
$$\{\pi^{\star}_k(1), \ldots, \pi^{\star}_k(\ell_k)\} \mid \boldsymbol{\zeta} \sim \text{Dir}\{\alpha^{\star}/\ell_k + n^{\star}_{k,1}, \ldots, \alpha^{\star}/\ell_k + n^{\star}_{k,\ell_k}\}.$$

Updating the random effects parameters

8: Let $\mathbf{r}^{(r)} = \{y_{i,\ell,t} - f_{x_1,\ldots,x_p}(t)\}_{i,\ell,t}$ be the random effects residuals. For $i = 1, \ldots, n$, sample the random effects curve parameters $\boldsymbol{\beta}^{(u)}_i$
$$p(\boldsymbol{\beta}^{(u)}_i \mid \boldsymbol{\zeta}) \sim \text{MVN}_K \left(\boldsymbol{\Sigma}^{\star}_i \mathbf{B}_i^{\intercal} \frac{\mathbf{r}^{(r)}_i}{\sigma^2_{\varepsilon}}, \boldsymbol{\Sigma}^{\star}_i \right),$$
where $\boldsymbol{\Sigma}^{\star}_i = \left(\sigma^{-2}_{\varepsilon} \mathbf{B}_i^{\intercal} \mathbf{B}_i + \sigma^{-2}_{u,s} \mathbf{P}_u + \sigma^{-2}_{u,a} \mathbf{I}_K \right)^{-1}$, $\mathbf{B}_i = \{B_k(t_i)\}_{t_i,k}$.

9: Sample the random effects smoothness parameter $\sigma_{u,s}$ using an M-H step from its full conditional
$$p(\sigma_{u,s} \mid \boldsymbol{\zeta}) \propto p_0(\sigma_{u,s}) \prod_{i=1}^{n} p(\boldsymbol{\beta}^{(u)}_i \mid \sigma_{u,s}, \sigma_{u,a}).$$

10: Sample the random effects scale parameter $\sigma_{u,a}$ using an M-H step from its full conditional
$$p(\sigma_{u,a} \mid \boldsymbol{\zeta}) \propto p_0(\sigma_{u,a}) \prod_{i=1}^{n} p(\boldsymbol{\beta}^{(u)}_i \mid \sigma_{u,s}, \sigma_{u,a}).$$

Updating the global variance parameter

11: Let $\mathbf{r} = \{y_{i,\ell,t} - f_{x_1,\ldots,x_p}(t) - u_i(t)\}_{i,\ell,t}$ be the residuals. Sample the error variance σ^2_{ε} as
$$\sigma^2_{\varepsilon} \mid \boldsymbol{\zeta} \sim \text{Inv-Ga}\{a_{\sigma} + nLT/2, b_{\sigma} + \mathbf{r}^{\intercal}\mathbf{r}/2\}.$$

4.4.2 Prior Hyper-parameters and MCMC Initializations

The fixed effects parameters of the longitudinal mixed effects model (4.1) are initialized at the maximum likelihood estimate for the spline coefficients of the simplified model with no smoothing and no predictors included. The random effects are initialized at zero.

The hyper-parameter for the half-Cauchy prior on the smoothing parameters is $s_\sigma = 1$. The $\mathrm{C}^+(0,1)$ distribution has its mode at zero and hence is capable of capturing strong smoothness but also has heavy tails and is thus also capable of capturing wiggly functions. The hyper-parameters for the inverse-Gamma prior on the residual variance are set at $a_\sigma = b_\sigma = 1$. The hyper-parameters on the Gamma prior for the mass of the Dirichlet distributions on the transition dynamics are $a_\alpha = b_\alpha = a_{\alpha^\star} = b_{\alpha^\star} = 1$, as recommended in Escobar and West (1995). Finally, the hyper-parameters for the Gamma prior on $\varphi_{j,k}$ are $a_{\varphi,j} = 5, b_{\varphi,j} = 1$.

4.4.3 Software, Runtime, etc.

The results reported in this chapter are all based on $7,500$ MCMC iterations with the initial $2,500$ iterations discarded as burn-in. The remaining samples were further thinned by an interval of 5. We programmed in `R` interfaced with `C++` through Rcpp (Eddelbuettel *et al.*, 2011) and RcppArmadillo (Eddelbuettel and Sanderson, 2014). The MCMC algorithm takes 10 minutes on a Macbook laptop with 8 Gb RAM for the synthetic examples. In all experiments, the posterior samples produced very stable estimates of the population and individual level parameters of interest. MCMC diagnostic checks were not indicative of any convergence or mixing issues.

4.5 Simulation Studies

In synthetic experiments, the proposed longitudinal framework achieved excellent empirical performance in recovering the true fixed and random effect curves and associated local cluster configurations from noisy subject level data. Figure 4.7 illustrates the scenario used in the simulation studies. We considered $T = 20$ time points $\{1, \ldots, T\}$. We generated $p = 10$ predictors, $x_1, x_2 \in \{1, 2\}$ and $x_3, \ldots, x_{10} \in \{1, 2, 3\}$. The total number of possible level combinations of $(x_1, \ldots, x_{10})$ across all time points to consider in a fully flexible but completely unstructured model would thus be $T \prod_{j=1}^{10} x_{j,\max} = 20 \times 2^2 \times 3^8 = 20 \times 26,244 = 524,880$. The true data generating mechanism is such that x_1 and x_3 are locally important whereas all other covariates are redundant at all time points. The fixed effects curves corresponding to the levels $\{1, 2\}$ and $\{3\}$ of x_3 are initially equal, then diverge at $t = 5$ and finally merge back at $t = 17$, conditional on $x_1 = 1$. The fixed effects curves corresponding to the levels $\{1\}$ and $\{2\}$ of x_1 are initially equal and then diverge at $t = 8$. We generated $n = 100$ individual specific curves with $L_{i,t} = 2$ repeated measurements at each time point. The residual variance was set at $\sigma^2_\varepsilon = 1$, whereas the variance and the smoothness of the random effects were $\sigma^2_{u,s} = 0.1$ and $\sigma^2_{u,a} = 2$, respectively.

As shown in Figure 4.6, our method correctly recovers x_1 and x_3 as the only significant predictors. In fact, the estimated number of groups $\ell_{j,k}$ associated with the other predictors consistently equals to one. The posterior probabilities also correctly estimate two groups for x_1 starting from $t = 8$ and two groups for x_3 starting from $t = 5$. Estimates of the fixed effects curves and a few individual level curves obtained by our method are shown in Figure 4.7. Our model estimates the fixed (left panel)

as well as the individual specific (right panel) effects very precisely by borrowing information whenever predictors are redundant or covariate levels are in the same cluster.

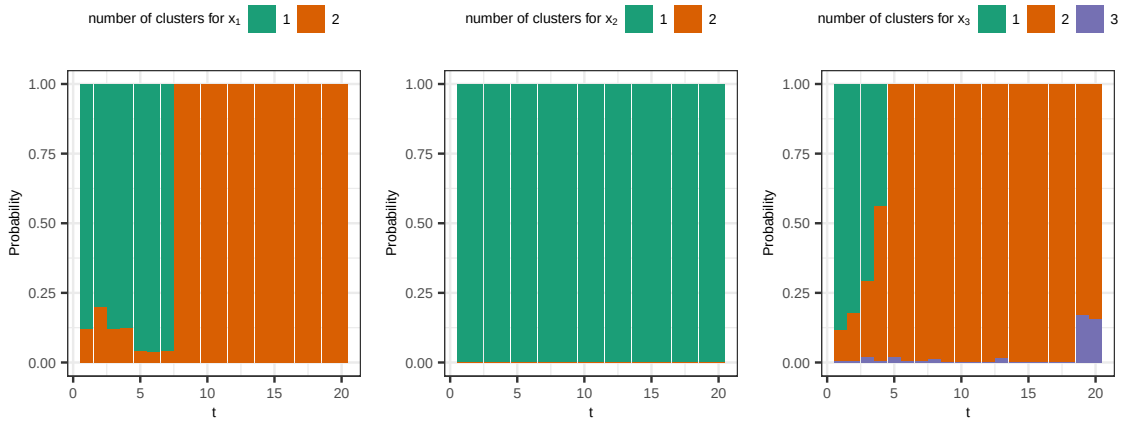


Figure 4.6: Results for synthetic data: The estimated posterior probabilities for the number of clusters of the predictors' levels over time for x_1, x_2 and x_3. The predictors x_1 and x_3 were locally important. The remaining predictors, namely $(x_2, x_4, \ldots, x_{10})$, including x_2 shown here, were never included in the model - their levels always formed a single cluster.

We compare the out-of-sample predictive performance of our proposed LFMM with state-of-the-art parametric and nonparametric regression alternatives. We focus particularly on BART models by fitting both the original BART (Chipman *et al.*, 2010) and the smooth BART (Linero and Yang, 2018) to the synthetic data sets. In addition, we apply a LASSO regression model, implemented using the function `glmnet` in `R`, independently at each time point. Figure 4.8 compares the out-of-sample predictive performance (left panel) and the coverage of the 95% prediction intervals (right panel) for the different methods for 100 simulated data sets with 75%-25% training-test splits. Remarkably, our proposed LFMM not only had substantially

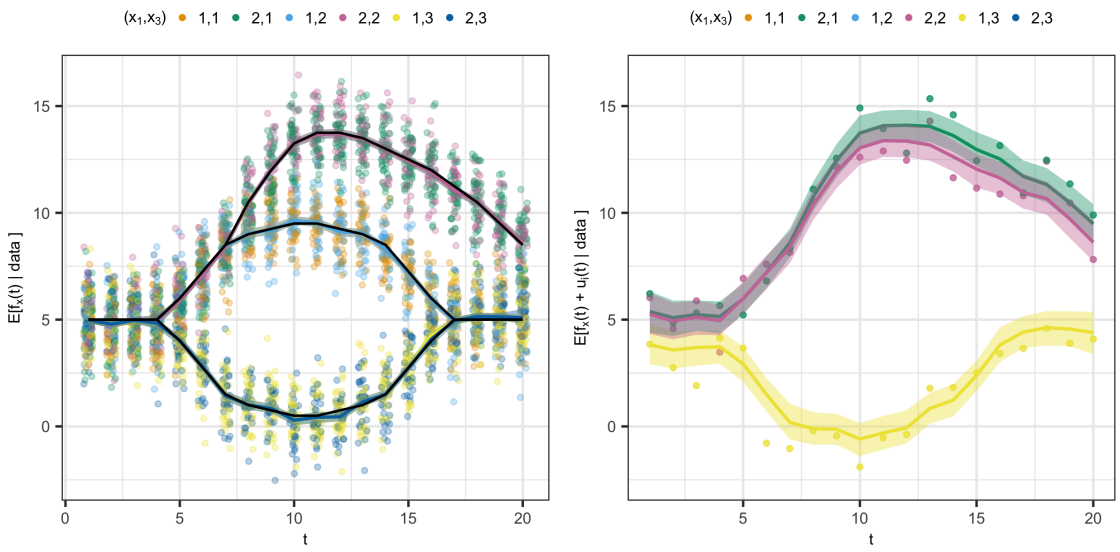


Figure 4.7: Results for synthetic data: Scenario with ten covariates $(x_1, \ldots, x_{10})$ where only (x_1, x_3) are locally important, as described in Section 4.5. Left panel: Estimated posterior means (colored lines) and 95% point wise credible intervals for the fixed effects, superimposed on slightly jittered response values $y_{i,\ell,t}$ for all combination of the levels of the significant predictors (x_1, x_3). The true fixed effects are superimposed (black lines). Right panel: Estimated posterior means (colored lines) and 95% point wise credible intervals for three individual specific curves, superimposed on the associated observed individual response values $y_{i,\ell,t}$. The figure here corresponds to the synthetic data set that produced the median root mean squared error.

smaller out-of-sample RMSEs, it actually performed uniformly better than all other approaches in all simulated data sets. Our method also produced prediction intervals with coverages probabilities close to the nominal rate.

As described in Section 4.2.1, our methodology can also recover main and interaction effects for the predictors included in the model. Figure 4.9 shows the

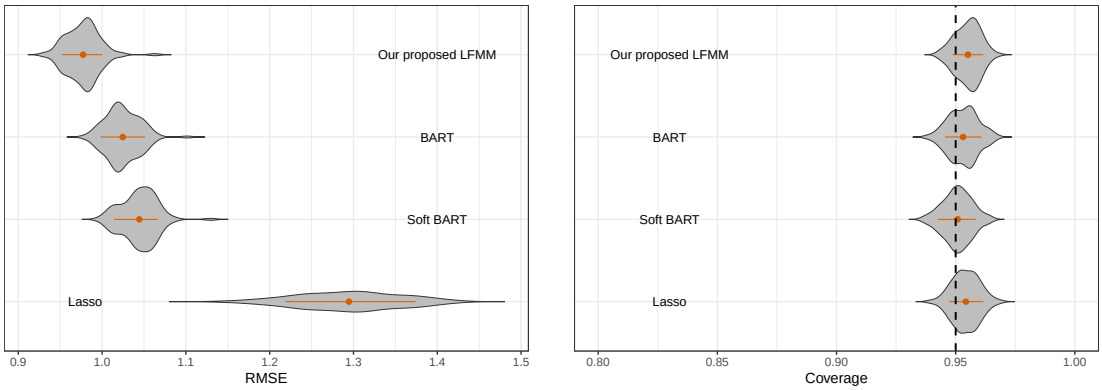


Figure 4.8: Results for synthetic data: The left panel shows the out-of-sample root mean squared error. The right panel shows the coverage of 95% prediction intervals. All measures reported are obtained over 100 75%-25% training-test splits. The red points represent the averages across simulations, whereas the red intervals represent the interquartile ranges across simulations.

true main and interaction effects for various levels and level combinations of x_1 and x_3, and the corresponding estimated posterior means and 95% point wise credible intervals.

4.6 Applications

In this section, we discuss the results of our method applied to five data sets, including one with time-varying predictors.

4.6.1 Progesterone Data

We describe here an application of our proposed approach to modeling progesterone data (Brumback and Rice, 1998; Nguyen and Gelfand, 2011) that record the logarithm of the progesterone levels of women during the course of their menstrual

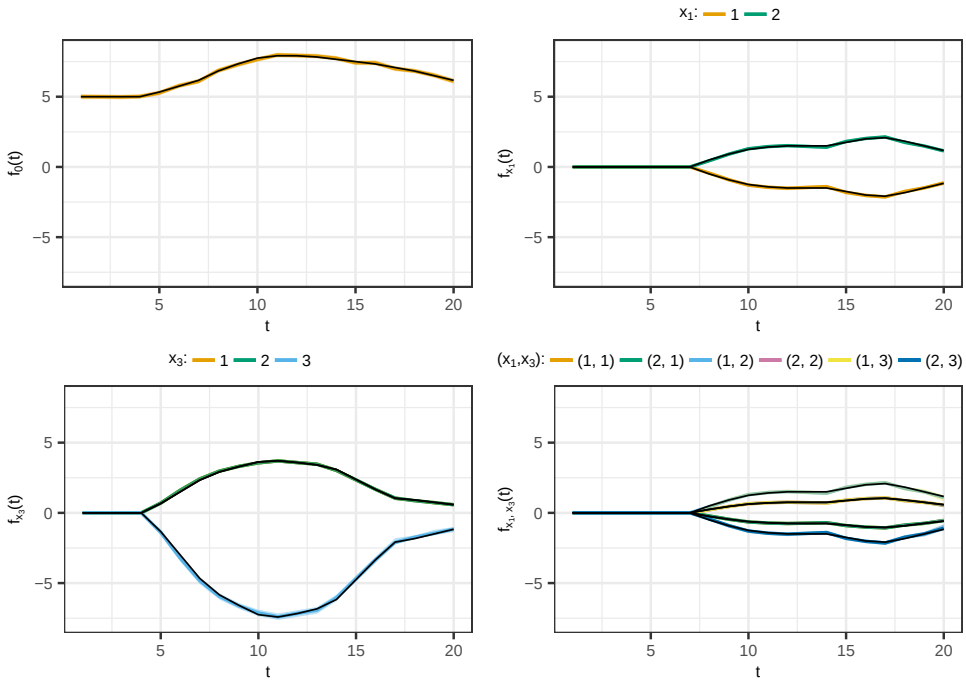


Figure 4.9: Results for synthetic data: Scenario with covariates $(x_1, \ldots, x_{10})$ with significant predictors (x_1, x_3) described in Section 4.5. Showing their true effects (black lines) and their estimated posterior means (colored lines). Clockwise from top left: overall mean; main effects of x_1; main effects of x_3; and interaction effects of (x_1, x_3). The figure here corresponds to the synthetic data set that produced the median root mean squared error.

cycles, measured by urinary hormone assay. Measurements of 51 female subjects occur during a monthly cycle ranging from -8 to 15 (8 days pre-ovulation to 15 days post-ovulation). There are a total of 91 cycles: the first 70 cycles belong to the non-conceptive group, the remaining 21 cycles belong to the conceptive group. The type of cycle is the single categorical predictor used in the analysis.

Figure 4.10 (left) shows the estimated posterior means and associated 95% point wise credible intervals for the group specific curves. The population level curves for conceptive and non-conceptive cycles are clustered together in the early

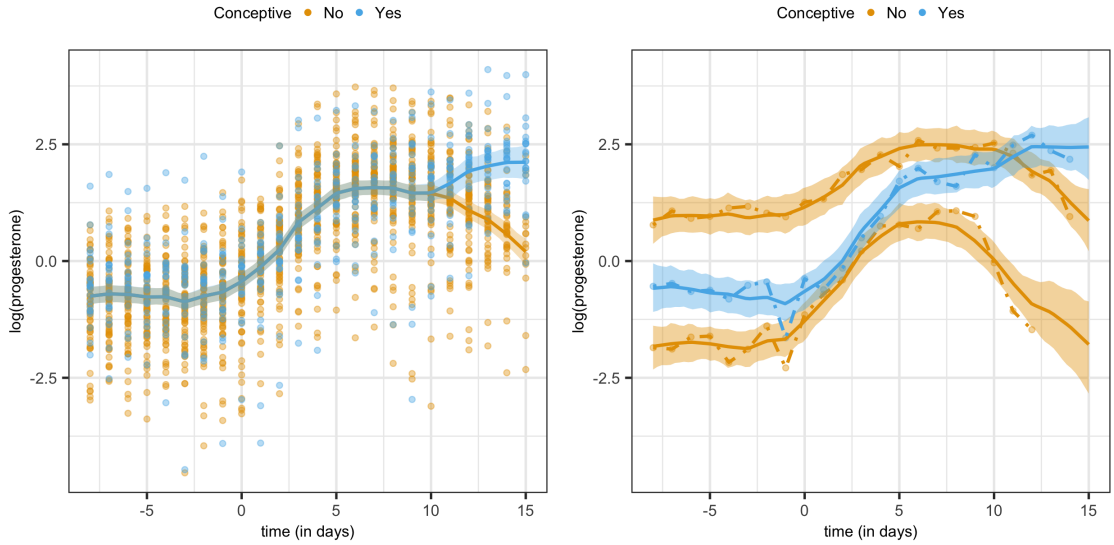


Figure 4.10: Results for the progesterone data: The left panel shows the estimated posterior means and 95% point wise credible intervals for the fixed effects curves, superimposed on slightly jittered response values $y_{i,\ell,t}$ for different levels of x. The right panel shows three examples of individual specific curves, their estimated posterior means (solid lines) and 95% point wise credible intervals, superimposed on the associated observed individual response values (dashed lines) $y_{i,\ell,t}$.

part of the cycle but become different in the late post ovulation period. In particular, the late conceptive cycles are associated with higher levels of progesterone. Global clustering methods would not allow clustering of the groups in the pre-ovulation period and would simply separate the two groups across all time points. Figure 4.10 (right) shows the estimated posterior means and associated 95% point wise credible intervals for the individual specific curves. These estimates show how our model can flexibly recover the individual level variations.

4.6.2 Health and Retirement Study Data

We analyze publicly available data from a longitudinal survey of US adults, the Health and Retirement Study (HRS). The HRS was established to assess the health implications of aging at both individual and population levels and has been fielded biennially years since 1992. Three categories of data - public, sensitive and restricted - can be accessed on the `HRS website` or, alternatively, via the `RAND HRS longitudinal file`. The HRS is sponsored by the National Institute on Aging and the University of Michigan and has previously been analyzed in Sonnega *et al.* (2014) and most recently in Deshpande *et al.* (2020).

The goal of the study is to understand how life course processes influence the trajectories of cognitive health. Therefore, we focus on predicting each subject's later-life cognitive function over time using life course socio-economic position (SEP) indicators. The $p = 13$ covariates include measures of SEP in childhood (SEP index), early adulthood (educational attainment), and later-life (household wealth) as well as measures of later-life mental and physical health (binary indicators of physical activity, diabetes, heart problems, high blood pressure, loneliness and stroke as well as BMI and depression index) and socio-demographic factors (race, gender). The size of the unstructured model $T \prod_{j=1}^{p} x_{j,\max} = 32 \times 580,608 = 18,579,456$ makes it impossible to estimate the parameters without adopting a dimensionality reduction approach. The outcome is cognitive function as measured by a series of listening and memory tests that the HRS used to construct a score ranging from 0 to 35. We restricted our analysis to subjects aged between 65 and 96 years with at least two cognitive scores recorded between 2000 and 2016. This resulted in a sample

of $n = 4,167$ subjects who were administered a total of $N = 27,820$ surveys, each individual being recorded either at even or at odd numbered ages but missing the intermediate values.

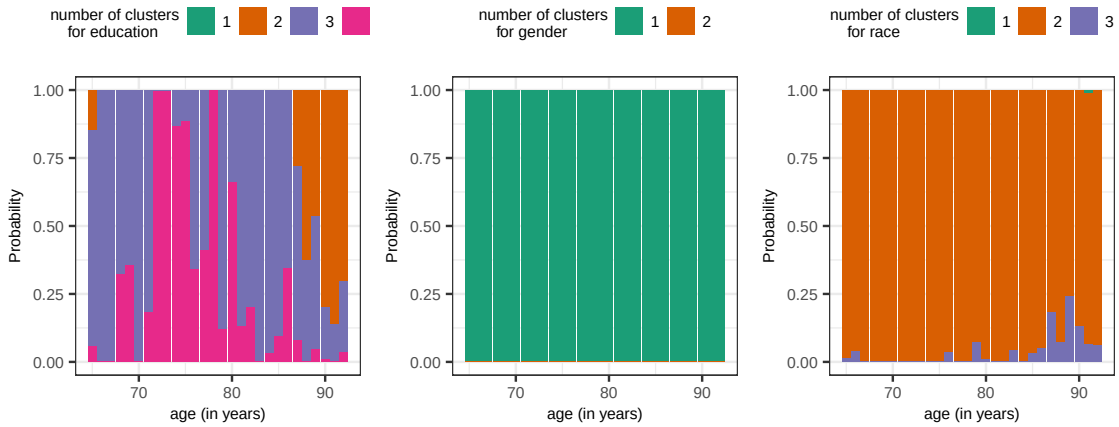


Figure 4.11: Results for the HRS: The estimated posterior probabilities for the number of clusters of the predictors' levels over time for $x_1 = education$, $x_2 = gender$, and $x_3 = race$. The predictors x_1 and x_3 were locally important. The remaining predictors, including x_2 shown here, were never included in the model since the number of clusters of their levels was always 1.

Figure 4.11 shows the posterior probabilities for the number of groups $\ell_{j,k}$ associated to three of the predictors (education, gender, race). The other predictors' levels were grouped together at each location k and therefore they did not affect the outcome. Figure 4.12 shows the effect of education and race, i.e., the two predictors that were selected by the model. These results highlight the importance of educational attainment due to its association with cognition. It appears that higher levels educational attainment are associated with higher cognitive function across adulthood. This confirms that socioeconomic position in early adulthood as measured by education can have later life effects on cognition. Conversely, it appears that

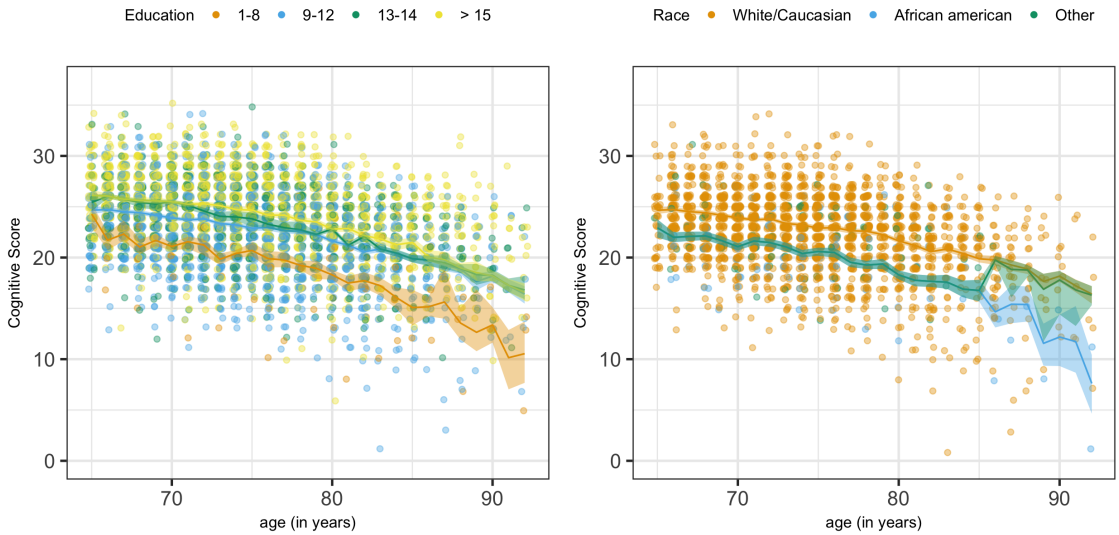


Figure 4.12: Results for the HRS: The left panel shows the estimated posterior means and 95% point wise credible intervals for the fixed effects curves corresponding to different education levels, superimposed on slightly jittered response values $y_{i,\ell,t}$. The right panel shows the estimated posterior means and 95% point wise credible intervals for the fixed effects curves corresponding to different races, superimposed on slightly jittered response values $y_{i,\ell,t}$.

the other SEP measures have no predictive effect on later-life cognition. In middle aged invididuals, three groups of educational attainment seem to differently affect the outcome: 1-8, 9-12, 13+. In old aged invididuals, instead, only two groups of educational attainment are significant: 1-8, and 9+. As far as race is concerned, it appears that after controlling for the other covariates in this study, white and non-white individuals have significant differences in cognitive scores during later-life. This finding also confirms the results in Deshpande *et al.* (2020), who estimated that white people's intercept parameter is larger than the one for other races, and is consistent with previous literature (Wilson *et al.*, 2015; Díaz-Venegas *et al.*, 2016). This result indicates that other factors that are unaccounted for (i.e., quality of ed-

ucation or literacy) are affecting the estimated cognitive scores for each race/ethnic group. Crucially, our model is able not only to flexibly estimate the cognitive score functions, but also to pool information across different covariate subgroups. Borrowing information across curves becomes especially important to estimate the cognitive score of older aged individuals due to the decrease in sample size.

4.6.3 Beat the Blues Data

We consider longitudinal data from a randomized clinical trial of an interactive multimedia program known as "Beat the Blues" which was designed to deliver cognitive behavioral therapy to depressed patients via a computer terminal. Patients with depression recruited in primary care were randomized to either the Beating the Blues program, or to "Treatment as Usual" (TAU), and they were followed up for a maximum of 4 visits. Other than the treatment indicator, the two additional predictors include dummy variables indicating if patients take anti-depressant drugs and if the length of the current episode of depression is less or more than six months. Thus, the size of the unstructured model $T\prod_{j=1}^{p} x_{j,\max} = 5 \times 2^3 = 40$ makes it hard to estimate the parameters with the small sample size $n = 380$, typical of a clinical trial. The measured outcome is the Beck Depression Inventory II (BDI), a popular depression screening instrument. The data is publicly available, for instance, via the `R` package `HSAUR2` (Hothorn and Everitt, 2014). The efficacy of computerized cognitive behavioral therapy was first detected in Proudfoot *et al.* (2003) via a linear mixed effects model but was not replicated in the randomized clinical trial of Gilbody *et al.* (2015).

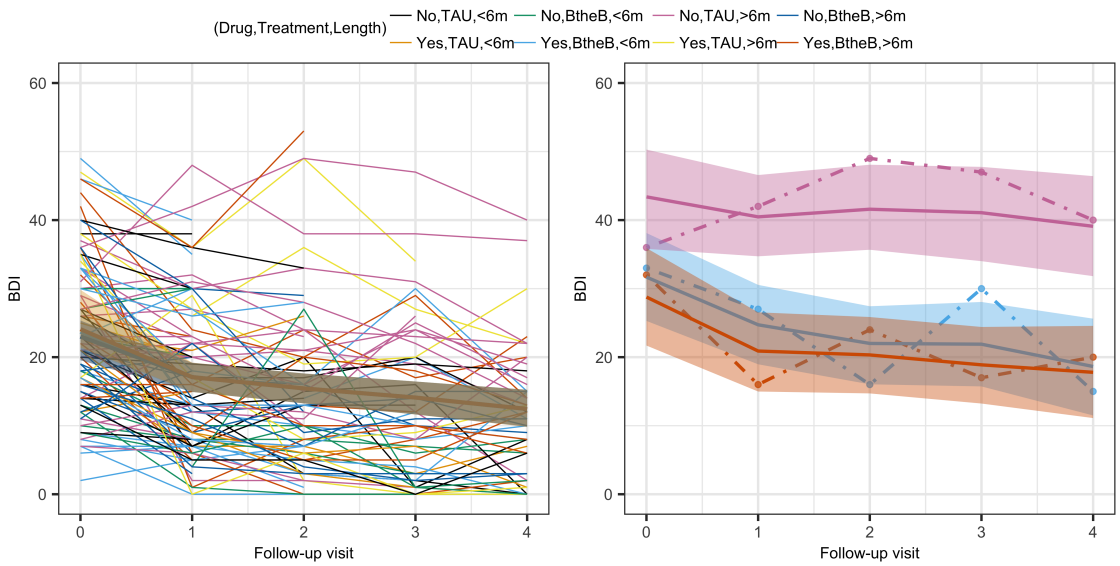


Figure 4.13: Results for the Beat the Blues data: The left panel shows the estimated posterior means and 95% point wise credible intervals for the fixed effects, superimposed on the observed trajectories $y_{i,\ell,t}$. The right panel shows the estimated posterior means and 95% point wise credible intervals for three individual specific curves, superimposed on the associated individual responses $y_{i,\ell,t}$.

Figure 4.13 (left panel) shows the estimated posterior means and associated 95% point wise credible intervals for the group specific curves for each of the eight possible combinations of the three categorical predictors. As illustrated, no significant differences were detected at any time point by our model. Thus, our model seems to confirm the conclusions of Gilbody *et al.* (2015). Figure 4.13 (right panel) shows the estimated posterior means and associated 95% point wise credible intervals for three individual specific curves.

4.6.4 Childhood Asthma Management Program (CAMP) Data

The CAMP study (Group, 2000) was a randomized clinical trial for children with asthma. We analyze a subsampled and anonymized version of the original data, publicly available at the National Institute of Health (NIH) `website`. We use this semi-synthetic data set for illustrative purposes alone, the conclusions must not be extended to the original study.

The trial's goal was to infer the long-term impact of three treatment assignments (Budesonide, Nedocromil, or placebo) on pulmonary function. A total of $n = 1,041$ children aged 5-12 years were enrolled and balanced across the treatment groups. We use one of the endpoints of the trial as the response variable, namely lung function as measured by the Forced Vital Capacity (FVC). Other predictors include the children's gender and ethnicity as well as dummy variables indicating if participants shared their house with pets and/or smokers. The size of the unstructured model is $T \prod_{j=1}^{p} x_{j,\max} = 16 \times 3 \times 2^3 \times 4 = 1,536$.

Figure 4.14 (left panel) shows the estimated posterior means and associated 95% point wise credible intervals for the fixed effects curves for each of the six possible paired combinations the predictors, namely treatment assignment and presence of smokers in the household. The only significant predictor in the model is the treatment assignment variable. In particular, participants assigned to Budesonide appear to have larger FVC. These differences, however, emerge only after the third visit and seem to become more pronounced as time progresses. Figure 4.14 (right panel) shows the estimated posterior means and associated 95% point wise credible intervals for three individual specific curves, exhibiting a high degree of heterogeneity around the

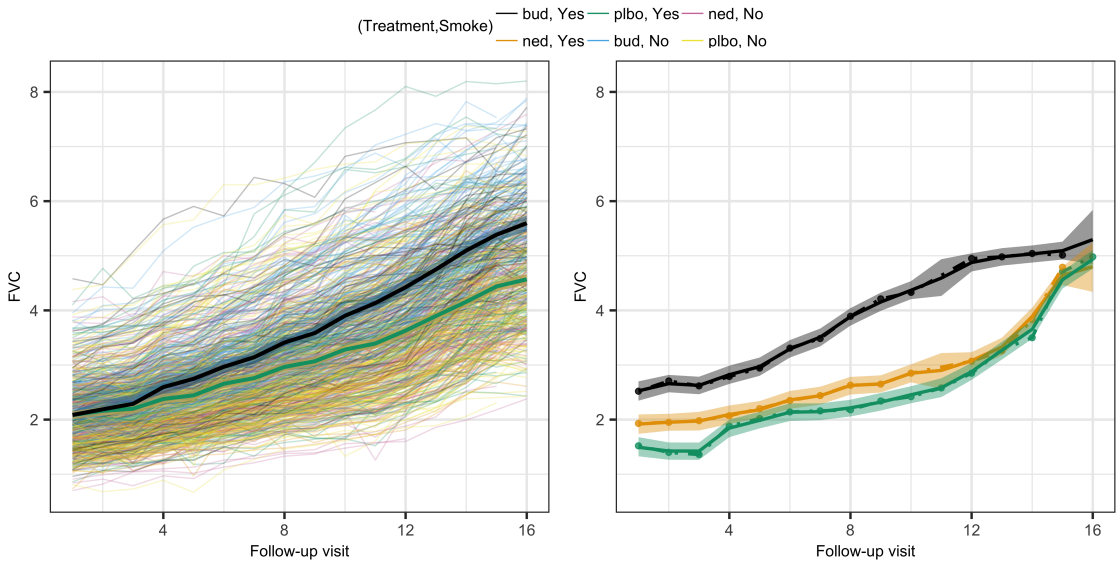


Figure 4.14: Results for the CAMP data: The left panel shows the estimated posterior means and 95% point wise credible intervals for the fixed effects, superimposed on the observed trajectories $y_{i,\ell,t}$. The right panel shows the estimated posterior means and 95% point wise credible intervals for three individual specific curves, superimposed on the associated individual responses $y_{i,\ell,t}$.

mean profiles.

4.6.5 National Longitudinal Survey of Youth Data

The national longitudinal survey of youth of 1997 (NLSY97, Moore *et al.*, 2000) is a longitudinal study that follows a nationally representative sample of the American youth born between 1980 and 1984 on various aspects of life. Participants enter the study between the ages of 12 and 16. Interviews were conducted annually from 1997 to 2011 and biennially since then. The NLSY97 collects information on respondents' labor market behavior and educational experiences. The survey also includes data on the participants' family backgrounds to help researchers assess the

impact of environmental factors on these labor statistics. We use a publicly available version of the data that can be found at the U.S. Bureau of Labor Statistics `website`.

We analyze a subsample of the original data consisting of a random 25% of the participants which resulted in $n = 2,120$ youths surveyed for a total of $16,188$ questionnaires. We use yearly income (in $\$10,000$ units) as the response variable and determine the effects of the socio-demographic variables on this outcome. The exogenous covariates include the participants' gender and ethnicity. The time-varying predictors are region, marital status and a dichotomous variable indicating if the participants live in urban or rural areas. The size of the unstructured model is $T \prod_{j=1}^{p} x_{j,\max} = 16 \times 2^2 \times 4^2 \times 5 = 5,120$.

Our analysis produced three significant predictors, namely gender, ethnicity and marital status. Displaying every level combination of these predictors is difficult, so we show only their main effects as defined in equation (4.5). Figure 4.15 shows the estimated posterior means and associated 95% point wise credible intervals for the overall mean (top left) and the predictors' main effects (other panels). The top left panel shows that incomes increase as a function of age on average across the entire sample. The top right panel shows a gender gap that becomes especially important between the ages of 21 and 26, with men earning up to $\$5,000$ more than women. A racial gap also appears to be significant, with African American participants having lower earnings, as illustrated in the bottom left panel. Finally, as illustrated in the bottom right panel, married couples seem to have higher incomes compared to single earners, perhaps an artifact of their joint filing of taxes.

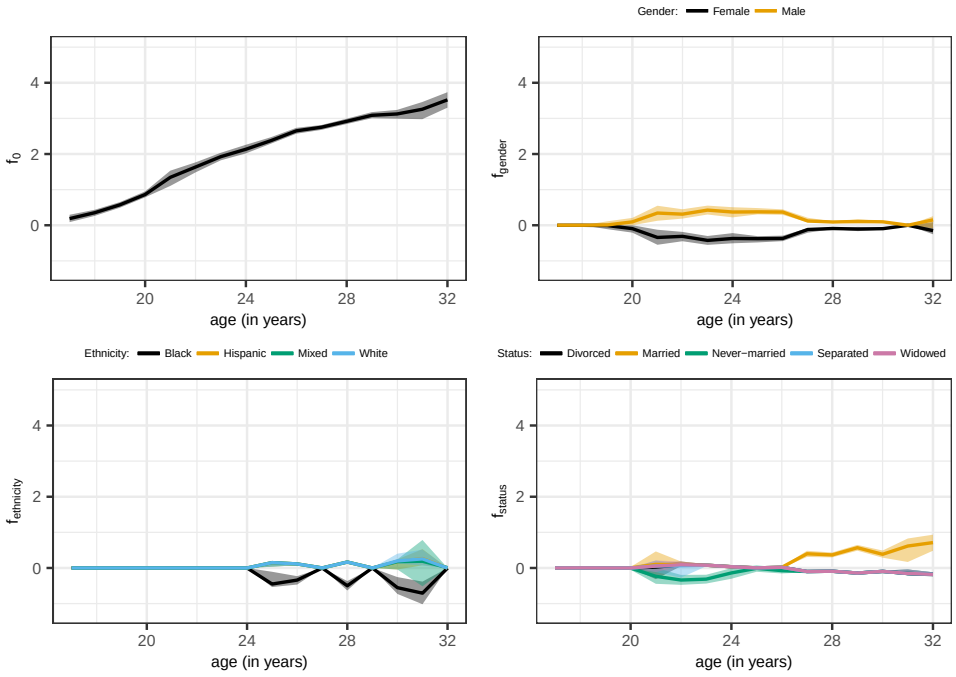


Figure 4.15: Results for the NLSY97 data: the significant predictors were gender, ethnicity and marital Status. Showing their estimated posterior means (colored lines) and 95% point wise credible intervals. Clockwise from top left: overall mean; main effects of gender; main effects of ethnicity; and main effects of marital status.

4.7 Discussion

In this chapter, we developed a flexible Bayesian semiparametric approach to longitudinal functional mixed models in the presence of categorical covariates. Building on novel fHMM infused mixtures of locally supported B-splines, our proposed method allows the fixed effects components to vary flexibly with the associated covariates, allowing potentially different sets of important covariates to be included in the model at different time points. The mechanism not only allows different sets of covariates to be included in the model at different time points but also allows the se-

lected predictors' influences to vary flexibly over time. Flexible time-varying additive random effects, modeled also by Markovian mixtures of B-splines, are used to capture subject specific heterogeneity. We established theoretical results on posterior consistency of the proposed method for both function estimation and variable selection. In simulation experiments, the method significantly outperformed the competitors. We illustrated the method's practical utility in real data applications.

The methodology presented here is highly generic and broadly adaptable to diverse other problems. While the focus of this work has been on dynamically varying longitudinal data models, the methodology could also be useful in static multiway mixed ANOVA designs. Methodological extensions we are pursuing as topics of separate research include dynamic partition models for observational units; spatial and spatiotemporal settings models for multivariate responses; principled approaches to accommodate continuous, ordinal and mixed type covariates; etc.

Chapter 5

Bayesian Nonparametric Bivariate Survival Regression for Current Status Data

This chapter considers Bayesian nonparametric inference for event time distributions based on current status data. We show that in the case of independent censoring conventional mixture priors, including the popular Dirichlet process mixture prior, lead to biologically uninterpretable results as they unnaturally skew the probability mass for the event times toward the extremes of the observed data. Simple assumptions on dependent censoring can fix the problem. We then extend the discussion to bivariate current status data with partial ordering of the two outcomes. In addition to dependent censoring, we also exploit some minimal known structure relating the two event times. We design a Markov chain Monte Carlo algorithm for posterior simulation. Applied to a recurrent infection study, the method provides novel insights into how symptoms-related hospital visits are affected by covariates.

5.1 Introduction

We develop Bayesian nonparametric survival regression for bivariate event times that are subject to a single censoring time. In particular, we consider bivariate current status data (Groeneboom and Wellner, 1992), referring to situations where

the only available information on each event time is whether or not it exceeds a monitoring time that is common to the two outcomes. Data of this type are often collected in studies on the prevalence of recurrent infectious diseases such as partner studies of HIV infections (Jewell and Shiboski, 1990), or in carcinogenicity testing when a tumor under investigation is occult (Dunson and Dinse, 2002). Wang and Ding (2000) show that the distribution for bivariate current status data is not identifiable using nonparametric maximum likelihood estimation. The goal of this chapter is twofold: first (Section 5.3), we propose a dependent censoring scheme that is useful for modeling univariate event time data; second (Section 5.4), we embed such dependent censoring within a flexible model that can identify the joint distribution of bivariate outcomes with the aid of weak structural assumptions.

Our goal is to develop a flexible model whose parameters have a biologically meaningful interpretation. Bayesian models are especially useful in such scenarios because of their ability to accommodate prior information. Nonparametric priors are often used to flexibly model a baseline survival function, usually completed with a parametric component that relates survival to a number of predictors. For example, Bayesian extensions of the proportional hazards (PH) model (Cox, 1972) have been proposed in Kalbfleisch (1978) and in Hjort (1990). Generalizations of the accelerated failure times (AFT) model (Buckley and James, 1979) based on a Dirichlet process prior appear in Christensen and Johnson (1988), Kuo and Mallick (1997), Kottas and Gelfand (2001), Hanson and Johnson (2004), or alternatively using Polya trees, for example in Hanson and Johnson (2002). In other cases the main inference target is the hazard function. Sparapani *et al.* (2016), for instance, construct nonparametric

survival regression using a Bayesian additive regression tree (BART) model (Chipman *et al.*, 2010) by adding time as an ordinal predictor to a BART-probit model for the hazard function.

In general, censored observations contribute limited information, via the distribution function or survival function as the corresponding factors of the joint likelihood. This becomes problematic in the case of current status data, as we shall demonstrate. Some proposals have been put forward to tackle these issues. In the case of univariate survival regression, generalizations of the PH model for current status data have been introduced in Cai *et al.* (2011), Wang *et al.* (2015) and in Huang (1996). Xue *et al.* (2004) propose a partly linear AFT model for univariate current status data. More similar to our approach, Wang and Ding (2000) model dependence between bivariate event times via a copula function. Dunson and Dinse (2002) use a Bayesian probit model with normal frailties to induce dependence among multivariate current status data. Nevertheless, there remains a gap in the literature concerning flexible nonparametric regression models for bivariate current status data under dependent censoring.

The motivating case study is inference for the Partner Notification Study (Golden *et al.*, 2005). The goal of the study is to understand the times of development of infection and symptoms for recurrent episodes of gonorrhea and/or chlamydial infections. The study design includes a single follow-up visit for each individual. During this visit the presence of symptoms and infection was recorded, leading to all censored data with shared censoring times for the two outcomes.

Let S denote the time of the onset of symptoms, I the time of infection,

and C the time of the hospital visit. Thus, four responses are possible: presence of both disease and symptoms ($I < C, S < C$), absence of both ($I > C, S > C$), absence of symptoms and presence of disease ($I < C, S > C$), and symptoms without disease ($I > C, S < C$). The latter can be explained by the fact that the surveyed symptoms are very generic and might also arise due to other underlying causes. This setup yields data that are bivariate in nature as two outcomes are registered. However, the censoring times, i.e. the hospital visit times, are restricted to a lower dimensional subspace, with a single follow-up visit to assess the presence of both symptoms and disease. Additional complexity arises from the partial ordering of the two outcomes: the infection time is a priori unlikely to follow the symptoms time. This can only occur when the symptoms arise due to other causes. Our model introduces features to reflect this consideration. We use a mixture model with one submodel being subject to an order constraint, representing symptoms due to the infection of interest, and another submodel without such constraint, allowing for symptoms due to other causes. While our discussion is motivated by a specific application, we note that similar data formats arise frequently in any study that involves data collection during follow-up visits. For example, doctors might record tumor recurrence using a CT scan and symptoms as reported by patients.

In the first part of this chapter, we demonstrate with simple examples the problems arising from the use of standard techniques with current status data. We then introduce structural assumptions that allow us to identify a meaningful distribution of the latent bivariate outcomes. We propose a Bayesian nonparametric (BNP) approach for modeling the joint distribution under these assumptions. An

important feature of BNP models is their large support, allowing us to approximate essentially arbitrary distributions (Ishwaran and James, 2001). To handle covariates, our approach is based on the dependent Dirichlet process (DDP) prior introduced by MacEachern (1999). See also the discussion in De Iorio *et al.* (2004) for the special case of categorical covariates.

Outline of the Chapter: The rest of this chapter is organized as follows. Section 5.2 describes the clinical study that motivates this article. Section 5.3 develops the proposed inference approach starting from a simple univariate case. Section 5.4 uses the univariate model as a building block for bivariate outcomes and outlines an MCMC strategy for estimation. Section 5.5 develops a Markov chain Monte Carlo (MCMC) algorithm for posterior computation. Section 5.6 presents the results of simulation studies. Section 5.7 presents the results of the proposed method applied to the Partner Notification Study. Section 5.8 finishes with concluding remarks.

5.2 The Partner Notification Study

The Partner Notification Study (Golden *et al.*, 2005) enrolled men and women who received a diagnosis of gonorrhea or genital chlamydia at most 14 days prior to enrollment. It was conducted in King County Seattle (Washington state, U.S.A.) from September 1998 to March 2003. Researchers contacted clinicians who diagnosed and treated the infections to seek permission to contact their patients. To minimize the likelihood of reinfection before randomization, patients who could not be contacted within 14 days after treatment were not eligible for the study, yielding a total of $n = 1864$ participants. The study was designed to gather current status

data of recurrent gonorrhea or chlamydial infection in patients 3 to 19 weeks after randomization to standard (control group, 933 individuals) or expedited partner therapy (intervention group, 931 individuals). The primary outcome was persistent or recurrent gonorrhea and/or chlamydial infection in the original participants within 90 days after enrollment, although actual follow up times varied considerably (19 to 161 days) due to both difficulty scheduling follow-up visits and anticipated hospitalizations due to symptoms. The issue of patient noncompliance is handled by our model via a dependent censoring mechanism. Sal y Rosas and Hughes (2011) previously analyzed data on infection times from the same study, explicitly allowing for outcome misclassification.

When visiting the hospital, two outcomes were recorded for each patient: presence of reinfection (I_i) and of symptoms (S_i). Thus, two latent event times (I_i, S_i) correspond to a common censoring time C_i, i.e. the time of the hospital visit. The data record for each patient C_i, and whether the patient has already experienced the infection $\Delta_{I_i} = \mathbb{1}(I_i < C_i)$ and some symptoms $\Delta_{S_i} = \mathbb{1}(S_i < C_i)$. While in general symptoms should follow the onset of infection, the definition of symptoms in this study is very generic and they might also be due to other causes. In the case $I_i < S_i$ it is impossible to tell whether symptoms are due to the disease of interest or any other cause, while when $I_i > S_i$ the symptoms are known to be due some other cause.

The recorded $n = 1832$ follow-up visits included patients reporting all four possible combinations of censoring for the two outcomes: $n_{00} = 1303$ patients did not experience symptoms and tested negative for the infection; $n_{10} = 121$ patients tested

positive for the infection but were not experiencing any symptoms (asymptomatic infections); $n_{01} = 325$ patients tested negative for the infection but were experiencing symptoms (due to other causes); $n_{11} = 83$ patients tested positive for the infection and were also experiencing symptoms (symptomatic infections).

Figure 5.1 shows two univariate nonparametric maximum likelihood estimates (MLE) (Groeneboom and Wellner, 1992) for the distributions of time to infection I_i and time to symptoms S_i, stratified by two covariates (gender and intervention) under the assumption of independent censoring. Female participants seem to experience symptoms sooner than men. The flat region of survival probability in the middle of the range of the observed data is typical for the nonparametric MLE and is clinically highly implausible. In Section 5.3 we show that the accumulation of probability mass toward the bounds of the observation range is a common issue when dealing with current status data. Moreover, these nonparametric MLE estimates represent marginal effects and do not take into account any correlation that is expected between the time to infection and time to symptoms.

5.3 Univariate Survival Analysis for Current Status Data

We introduce a Bayesian nonparametric (BNP) modeling strategy for current status data, first in a simple univariate case. We show that the nonparametric MLE for current status data under independent censoring has an undesirable feature that makes it biologically uninterpretable. More specifically, most of the probability mass is accumulated toward the extremes of the data range.

Let S_i represent the latent event time for patient i, Δ_i be a censoring indicator

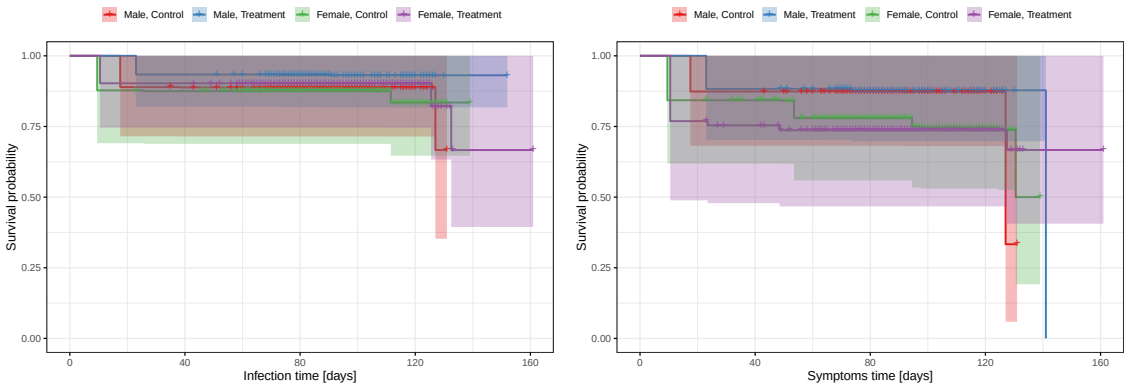


Figure 5.1: Nonparametric MLE for infection times (left panel) and time until symptoms (right panel), stratified by the binary covariates *gender* and *treatment* fixing *age* to the average age in the sample. Shaded areas represent pointwise 95% confidence intervals.

with $\Delta_i = 1$ if the event has been detected and $\Delta_i = 0$ otherwise, and let C_i denote the censoring time. That is, when $\Delta_i = 1$, then $S_i \leq C_i$ (left censored), otherwise $S_i > C_i$ (right censored). We want to infer the unknown density $f_S(s)$ based on only the observed censoring times and indicators $(C_i, \Delta_i), i = 1, \ldots, n$.

5.3.1 Limitations of the Maximum Likelihood Estimator

We show that under moderate sample sizes the nonparametric MLE under independent censoring does not provide meaningful estimates of the event time distribution for current status data. Without loss of generality, we assume that the censoring times are ordered, $C_i \leq C_{i-1}$, and that $\Delta_1 = 1, \Delta_n = 0$. Define $A = \{i > 1 \text{ s.t. } \Delta_i = 1, \Delta_{i-1} = 0\} \cup \{1\}$ as the set of indices of left censored observations immediately following a right censored observation, i.e. the set of indices of the pairs $(\Delta_{i-1}, \Delta_i) = (0, 1)$. Next, let $J = |A|$ and $\mathbf{C}^\star = (C_1^\star, \ldots, C_J^\star) = (C_i, i \in A)$ denote the corresponding censoring times. See Figure 5.2 for an illustration.

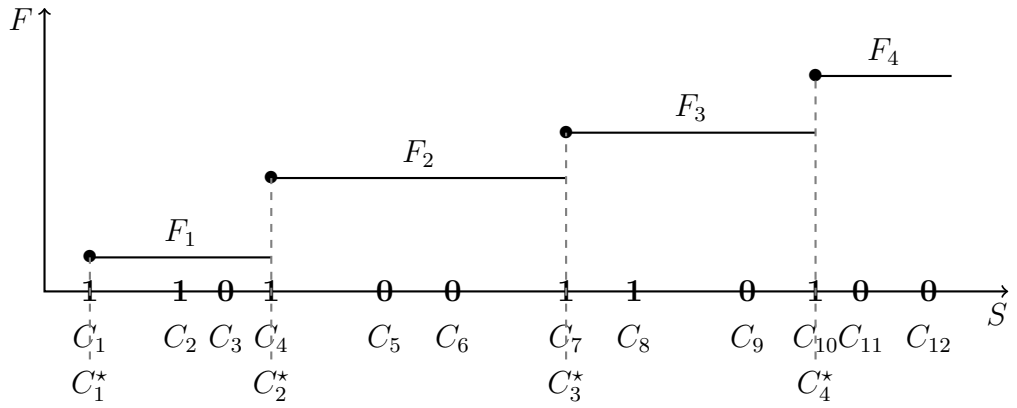


Figure 5.2: An example with $n = 12$ latent event times. The set of support points is $A = \{1, 4, 7, 10\}$. On the x-axis, 0 and 1 indicate the values of Δ_i.

Let $C^\star_{J+1}$ denote any point to the right of the last right censored observation. The times $\mathbf{C}^\star \cup \{C^\star_{J+1}\}$ are the only points where probability mass can accumulate under the nonparametric MLE. In other words, the support of a discrete nonparametric density estimate for the latent event times can have probability mass only at the left censoring times. More specifically, the support of the MLE is restricted to C_i's corresponding to (i) the left censored observation in every "01" pair, (ii) the first left censored observation, and (iii) any point to the right of the last right censored observation. To see this, write the unknown distribution $f_S(\cdot)$ of the latent times S_i as a discrete probability measure with atoms at the $C^\star_j$, i.e.

$$f_S(s) = \sum_{j=1}^{J+1} p_j \delta_{C^\star_j}. \tag{5.1}$$

We denote with $F_j = \sum_{k \le j} p_k$ the cumulative density function (c.d.f.) and with $\bar{F}_j = 1 - F_j$ the survival function at the supporting point $C^\star_j$. To see that the nonparametric MLE for $f_S(s)$ can only have support on the set $\mathbf{C}^\star$, assume that

$f_S(s)$ were to include any additional probability mass p at $C_i \neq C_j^\star, j = 1, \ldots, J$. Let $j^\star = \max_j\{C_j^\star < C_i\}$ and $j' = \min_j\{C_j^\star > C_i\}$ denote the point mass in $\mathbf{C}^\star$ closest to C_i from the left and from the right, respectively. Then, if $\Delta_i = 1$ one could move the probability mass p to $C_{j^\star}^\star$, and if $\Delta_i = 0$ one could move the probability mass p to $C_{j'}^\star$. Either would leave the likelihood function unchanged.

Groeneboom and Wellner (1992) introduce a simple EM algorithm to estimate the unknown c.d.f for the latent times under the independent censoring assumption. Let $l_j = \#\{C_i \text{ s.t. } \Delta_i = 1, C_j^\star \leq C_i < C_{j+1}^\star\}$ and $r_j = \#\{C_i \text{ s.t. } \Delta_i = 0, C_j^\star < C_i \leq C_{j+1}^\star\}$ denote the runs of left and right censored observations, respectively. Let $\mathbf{Y} = \{(C_i, \Delta_i)\}_{i=1}^n$ denote the data and $\mathbf{p} = \{p_j\}_{j=1}^{J+1}$ denote the parameters. The log-likelihood function under model (5.1) is

$$\begin{aligned}\ell(\mathbf{p}; \mathbf{Y}) &= \sum_{i=1}^{n}\{\delta_1(\Delta_i) \cdot \log F(C_i) + \delta_0(\Delta_i) \cdot \log \bar{F}(C_i)\} \\ &= \sum_{j=1}^{J}\{l_j \log F_j + r_j \log \bar{F}_j\}.\end{aligned}$$

If instead we knew the latent times $\mathbf{z} = \{S_i\}_{i=1}^n$, we could use the full data log-likelihood $\ell(\mathbf{p}, \mathbf{z}) = \sum_{j=1}^J n_j \log(p_j)$ where $n_j = \#\{S_i = C_j^\star\}$. The expectation of this full data log-likelihood with respect to $\mathbf{z}$ involves only $\mathbb{E}(n_j \mid \mathbf{p})$. This motivates an easy Expectation Maximization (EM) algorithm, illustrated below.

We illustrate the algorithm on simulated data with $n = 200$ latent times generated from a mixture of three normal distributions with weights $\boldsymbol{\pi} = (0.4, 0.2, 0.4)^\intercal$, locations $\boldsymbol{\mu} = (20, 40, 60)^\intercal$ and scale parameters $\boldsymbol{\sigma}^2 = (25, 25, 25)^\intercal$. The censoring times C_i were simulated according to model (5.2), defined below. As shown in Figure

Algorithm 2 (Expectation Maximization)

E-step

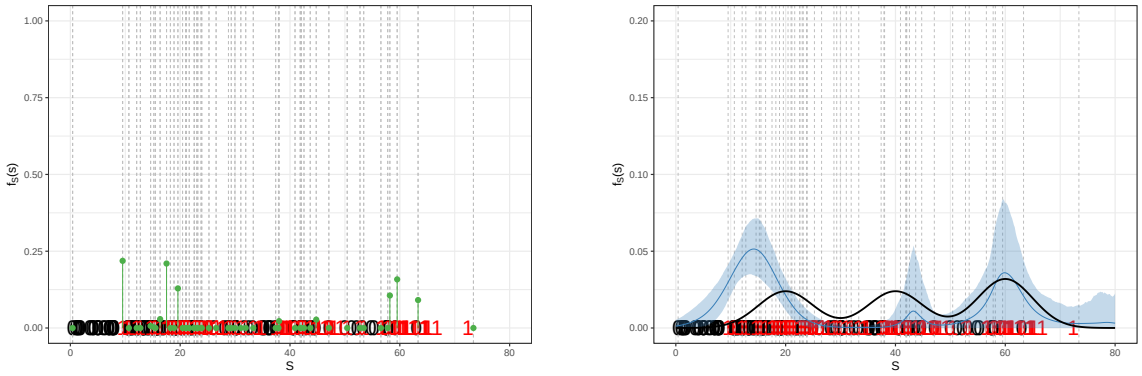
1: For given $\mathbf{p} = \mathbf{p}^{(t)}$, evaluate the expectation of the n_j's w.r.t. the latent censored event times. This involves distributing l_j to all $C^\star_{j'}$, $j' \leq j$ with weights $p_{j'}/F_j$; and r_j to all $C^\star_{j'}$, $j' > j$ with weights $p_{j'}/\bar{F}_j$, i.e.

$$\tilde{n}_j = \mathbb{E}_{\mathbf{s}}[n_j \mid \mathbf{p}] = \sum_{h \leq j} l_h p_j / F_h + \sum_{h > j} r_h p_j / \bar{F}_h.$$

M-step

2: Replacing the unknown n_j's with their expectations $\tilde{n}_j$ makes the maximization (w.r.t $\mathbf{p}$) of the expected (w.r.t $\mathbf{z}$) full data log-likelihood $\ell(\mathbf{p}, \mathbf{z})$ straightforward, yielding the update

$$p_j^{(t+1)} = \tilde{n}_j / n.$$



(a) Green vertical pins represent the non-parametric MLE estimate for the point masses obtained via the EM algorithm.

(b) In blue, posterior mean (and shaded pointwise 95% credible intervals) for a simple mixture of $K = 3$ normal distributions. In black, the simulation truth.

Figure 5.3: Simulated data. Right and left censoring times are represented by black "0" and red "1", respectively, on the x-axis. Vertical dashed lines represent the possible support points for $f_S(s)$.

5.3a, despite a large number of support points $\mathbf{C}^\star$, in this simulation study most of the probability mass under the unconstrained MLE accumulates close to the bounds of the range of the data. One might conjecture that the issue is caused by the excessively flexible nature of the unconstrained MLE. However, even parametric models fail to capture the underlying distribution of the latent times. For comparison, we carried out inference using a mixture of $K = 3$ Gaussian distributions for the latent times S, matching the nature of the actual simulation truth. In Figure 5.3b, we show the posterior mean for the unknown event time distribution under this model when fitted to the current status data in the simulation study. The posterior estimated distribution still allocates most probability mass toward the extremes of the data and misses the central peak, despite using an analysis model that matched the actual simulation truth.

5.3.2 A Bayesian Nonparametric Model

We introduce some assumptions to address the issues described in the previous section. In short, we regularize the model by (i) explicitly modeling the dependence between censoring times and latent event times, and (ii) introducing prior shrinkage with a flexible nonparametric Bayesian prior.

Knowledge about dependent censoring allows us to gain some information on $f_S(\cdot)$ from the censoring times. For example, in the motivating case study it is expected that patients seek help shortly after they experience symptoms. This information can be incorporated in the model in many ways. For our specific application, we assume that the censoring times C_i's arise from a race between a return

by schedule versus a return driven by the onset of symptoms, as

$$C_i \mid S_i, \lambda = \min\{S_i + \text{Exp}(\lambda); \text{Unif}(A, B)\}, \tag{5.2}$$

where A and B represent the range of the observation window, and $\text{Exp}(\lambda)$ and $\text{Unif}(A, B)$ refer to random variables with the respective distribution. In other words, the visit time to the hospital can either occur uniformly in the observation range (visit by protocol) or it can closely follow the symptoms onset (visit prompted by symptoms). The resulting distribution is easily evaluated.

Lemma 1. *The p.d.f. of the conditional distribution of censoring times given the event times is given by*

$$f_{C|S}(c \mid s) = \frac{\mathbb{1}\{c \leq s\}}{B - A} + \frac{\mathbb{1}\{c > s\}}{B - A} e^{-\lambda(c-s)}\{1 + \lambda(B - c)\}.$$

Proof. *Recall that*

$$\{C \mid S = s\} = \min\{s + \mathit{Exp}(\lambda); \mathit{Unif}(A, B)\}.$$

Then, the inverse cumulative density function for the conditional distribution of censoring times given the latent times is given by the survival function

$$\begin{aligned}
\bar{F}_{C|S}(c) &= \mathbb{P}[\min\{s + \mathit{Exp}(\lambda); \mathit{Unif}(A, B)\} > c] \\
&= \mathbb{P}[s + \mathit{Exp}(\lambda) > c; \mathit{Unif}(A, B) > c] \\
&= \mathbb{P}[s + \mathit{Exp}(\lambda) > c] \cdot \mathbb{P}[\mathit{Unif}(A, B) > c] \\
&= \frac{B - c}{B - A} \mathbb{1}_{(A,B)}(c) \left[\mathbb{1}\{c \leq s\} + e^{-\lambda(c-s)} \mathbb{1}\{c > s\}\right].
\end{aligned}$$

The condition $\mathbb{1}_{(A,B)}(c)$ will be considered to be always true, and hence omitted, in the following. This is assured by choosing A and B such that they cover the observation

range. Therefore,

$$F_{C|S}(c) = \left(1 - \frac{B-c}{B-A}\right) \mathbb{1}\{c \leq s\} + \left(1 - \frac{B-c}{B-A} e^{-\lambda(c-s)}\right) \mathbb{1}\{c > s\},$$

and by differentiation we get

$$f_{C|S}(c) = \frac{1}{B-A} \mathbb{1}\{c \leq s\} + \frac{e^{-\lambda(c-s)}}{B-A} \{1 + \lambda(B-c)\} \mathbb{1}\{c > s\}.$$

■

The regularization induced by the dependent censoring mechanism in (5.2) yields more intepretable inference, but some issues remain. Figure 5.4 shows the nonparametric density estimate for such a model under dependent censoring, and it highlights that inference still fails to recover the simulation truth. Two important features that are missing from this model are prior smoothing for the distribution of the latent event times as well as borrowing of information within homogeneous patient subpopulations.

Motivated by the described limitations we specify a Bayesian nonparametric prior for the latent event times. Relaxing parametric assumptions allows for greater modeling flexibility, robustness against misspecification of a parametric statistical model and, as a result, more honest uncertainty assessment than under a parametric model. At the same time, prior smoothing and shrinkage result in more realistic and clinically meaningful estimates compared to a nonparametric MLE. In addition, a BNP model can allow to accommodate heterogeneous patient populations, for example using a Dirichlet process (DP) mixture model.

Let f denote the distribution of the variable of interest (in our case the event times). A DP mixture model assumes $f(y) = \int k(y \mid \theta)\, dH(\theta)$ with $H \sim DP(M, H_0)$, where DP indicates a DP prior with total mass α and base measure H_0. See Chapter 2.1 for a more thorough review of the DP and DP mixtures. For later reference we note that $H = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}$ is a.s. discrete with $\theta_h \overset{\text{iid}}{\sim} H_0$, and a stick breaking prior (Sethuraman, 1994) for the weights $\pi_h = q_h \prod_{\ell<h}(1 - q_\ell)$ with $q_h \overset{\text{iid}}{\sim} \text{Beta}(1, M)$. Two natural choices of sampling models $k(y \mid \theta)$ for survival data are the log normal and the Weibull families. In applications with event times close to 0, it can be convenient to first log transform the data and then use normal kernels, i.e. use log normal kernels. In many instances, however, a mixture of normals may suffice (Lo, 1984) and is often preferred.

The BNP-CS model

The resulting model can be summarized as

$$\begin{aligned} C_i \mid S_i, \lambda &= \min\{S_i + \text{Exp}(\lambda); \text{Unif}(A, B)\} \\ S_i \mid H &\sim \int \text{N}(S_i \mid \mu, \sigma^2) dH(\mu, \sigma^2), \quad H \sim \text{DP}(M, H_0). \end{aligned} \tag{5.3}$$

The model is completed with hyperpriors

$$H_0 = \text{N}(\mu_k \mid \mu_0, \sigma_k^2/\kappa_0) \times \text{IG}(\sigma_k^2 \mid a_\sigma, b_\sigma),$$

$M \sim \text{Gamma}(a_M, b_M)$ and $\lambda \sim \text{Gamma}(a_\lambda, b_\lambda)$. We refer to (5.3) as BNP for current status (BNP-CS) model, with the name implying that alternative BNP priors other than the DPM (see, e.g. Müller *et al.*, 2015) could be used if desired. Using the stick-

breaking construction of the DP, the second line of model (5.3) can be rewritten as

$$S_i \mid \{\mu_k, \sigma_k^2, \pi_k\}_{k=1}^{+\infty} \sim \sum_{k=1}^{+\infty} \pi_k \mathrm{N}(S_i \mid \mu_k, \sigma_k^2)$$

with $(\mu_k, \sigma_k^2) \sim H_0$, i.i.d., and $\boldsymbol{\pi} \sim \mathrm{SB}(M)$, where $\mathrm{SB}(M)$ denotes the stick-breaking construction for the weights, with concentration parameter M. In our implementation, we also use priors on the hyperparameters $\mu_0, \kappa_0, b_\sigma$.

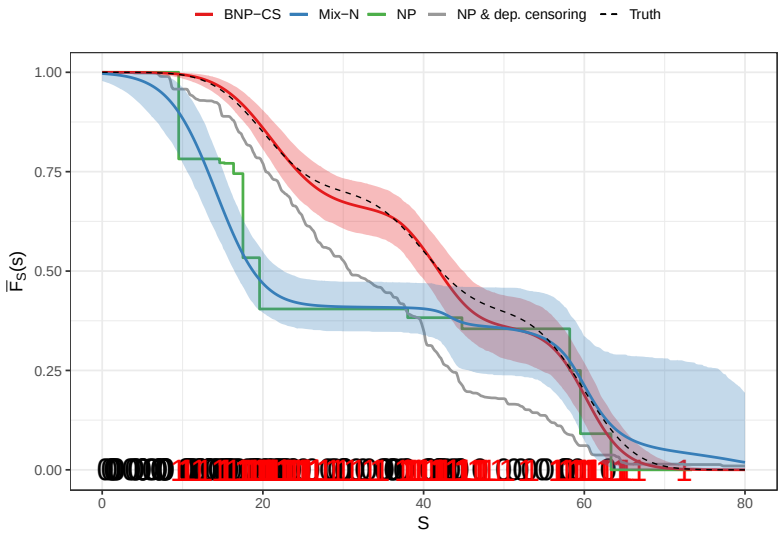


Figure 5.4: Simulated data: Right and left censoring times are represented by "0" and "1", respectively, on the x-axis. The green step function shows an estimate of the survival function under the nonparametric MLE using independent censoring. The gray step function shows an estimate of the survival function under the nonparametric MLE using dependent censoring. The blue curve shows an estimate of the survival function under a mixture of normals model (note that the simulation truth is in the same parametric family). The red curve shows an estimate of the survival function under the proposed model. Shaded areas represent pointwise 95% credible intervals for the estimated survival functions. The black dashed line represents the simulation truth.

Inference under the BNP-CS model for the same data used in the illustra-

tion of Section 5.3.1 recovers the underlying truth better than inference under the model with independent censoring. Figure 5.4 shows the survival function estimated under (i) an unconstrained nonparametric model, (ii) a nonparametric model under dependent censoring, (iii) a mixture of $K = 3$ normal distributions with independent censoring, and (iv) the proposed nonparametric model with dependent censoring. Although the model under (iii) matches the simulation truth, inference under models (i) - (iii) fails to recover a meaningful estimate, while inference under (iv) successfully exploits the information that is contained in the observed C_i.

5.4 Bivariate Survival Regression for Partially Ordered Current Status Data

5.4.1 A Bivariate Event Time Model

We now use the BNP-CS model (5.3) as a building block for bivariate outcomes. Beyond the already discussed dependence of S_i and C_i, we add some more structure based on prior knowledge of the underlying process. Without additional assumptions, the joint distribution for bivariate current status data is not likelihood identifiable, in general (Wang and Ding, 2000).

To see that this is the case, let $F_I = P(I_i \leq C_i) = P(\Delta_{I_i} = 1)$, $F_S = P(S_i \leq C_i) = P(\Delta_{S_i} = 1)$, and $F_{IS} = P(I_i \leq C_i, S_i \leq C_i) = P(\Delta_{I_i} = \Delta_{S_i} = 1)$. Note that F_I, etc. are functions of C_i. Assuming independent censoring the joint likelihood function for bivariate current status data $\mathbf{Y} = (\Delta_{I_i}, \Delta_{S_i}, C_i, i = 1, \ldots, n)$ is then

$$\prod_i \left\{ F_{IS}^{\Delta_I \Delta_S} (F_I - F_{IS})^{\Delta_I (1-\Delta_S)} (F_S - F_{IS})^{(1-\Delta_I)\Delta_S} (1 - F_S - F_I + F_{IS})^{(1-\Delta_I)(1-\Delta_S)} \right\}, \tag{5.4}$$

where, for ease of notation, we suppressed the i index in Δ_I and Δ_S. Only the three univariate distributions F_I, F_S and F_{IS} are likelihood identifiable. To achieve inference on the joint distribution of (I, S) we can therefore either (i) estimate the joint distribution under parametric or semiparametric assumptions, or (ii) build the joint model from the two identifiable marginal distributions and a particular choice for their dependence structure. Our approach follows mainly the latter strategy.

In words, we assume that symptoms can arise either due to the infection of interest, or due to other causes. In the former case, we assume a parametric model for the lag time $L = S - I$ between infection time I and onset of symptoms S. In the latter case, we assume independence between I and S. That is, we model the bivariate event time distribution $f_{IS}(I, S)$ as a mixture model in which one of the two components is subject to the order constraint $I < S$, i.e.

$$f_{IS}(I, S) = w f^{\star}_{IS}(I, S) + (1 - w) f'_{IS}(I, S) \tag{5.5}$$

where $f'_{IS}(I, S)$ is subject to $I < S$, whereas $f^{\star}_{IS}(I, S)$ is not. Therefore, $f^{\star}_{IS}(I, S)$ can be interpreted as the distribution of (I, S) for a patient with symptoms "due to other causes". Figure 5.5 shows the support of the two components of the mixture as well as the support for the latent times corresponding to the four possible censoring indicators, i.e. factors in (5.4).

We add two main assumptions to introduce more structure in (5.5), which will eventually facilitate inference: (i) under $f^{\star}_{IS}(I, S)$, the time to symptoms (due to other causes) and time to infection are independent; (ii) under $f'_{IS}(I, S)$, the latency time $L = S - I$ and the time to infection are independent. Here L is the delay from

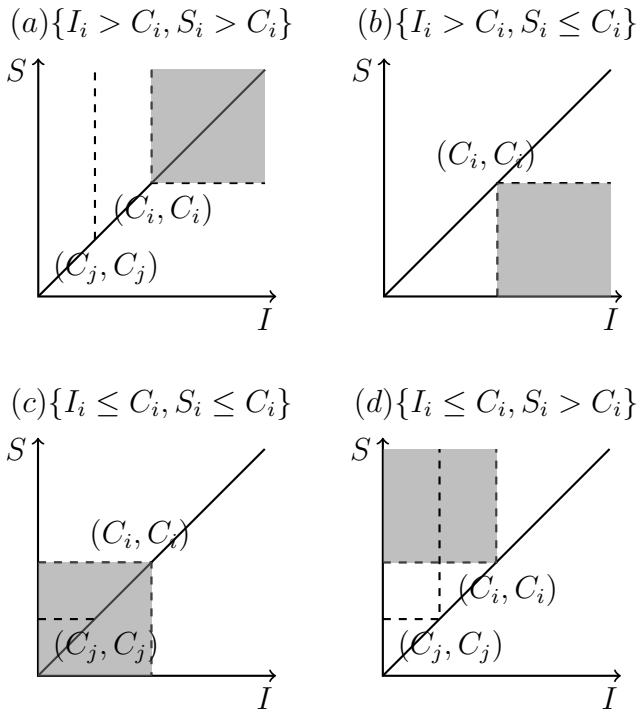


Figure 5.5: Support for the latent times $I > 0, S > 0$, corresponding to the four cases. The gray quadrants represent the support for the latent times corresponding to the observed censoring times (C_i, C_i) under $f^{\star}_{IS}(I, S)$. The area with red horizontal lines represents the support for the latent times under $f'_{IS}(I, S)$.

the onset of illness to the development of symptoms. The assumed marginal $f_I(\cdot)$ on I is shared by both, $f^{\star}_{IS}$ and f'_{IS}. Thus, model (5.5) becomes

$$f_{IS}(I, S) = w f_I(I) f^{\star}_S(S) + (1 - w) f_I(I) f_L(S - I). \tag{5.6}$$

Finally, note that by introducing in (5.6) dependence between S and I, we implicitly also introduce dependence between I and C through (5.3), thus regularizing inference on both f_I and f_S. For later reference we note that sampling $(I_i, S_i) \sim f_{IS}$ can be equivalently written as a hierarchical model with latent indicators, say v_i, with

$p(v_i = 1) = w$ and

$$(I_i, S_i \mid v_i) \sim \begin{cases} f_I(I) f_S^\star(S) & \text{if } v_i = 1 \\ f_I(I) f_L(S - I) & \text{if } v_i = 0. \end{cases} \tag{5.7}$$

The second component in (5.6) includes the constraint $I < S$ as a positivity constraint on the latency time $L > 0$. Recent approaches to deal with hard constraints use relaxation methods that replace the hard constraint with priors that penalize departures outside of the constraint subspace (Duan *et al.*, 2020). Alternatively, Patra and Dunson (2018) developed methodology that uses unconstrained inference and then projects the posterior draws onto the constrained subspace. In our model, assigning positive support to the reparametrized variable L automatically ensures the required order constraint $I < S$. In the following, we will use $L \mid \lambda_L \sim \text{Exp}(\lambda_L)$. As a consequence, under $f'_{IS}(I, S) = f_I(I) f_L(S - I)$, time to symptoms and time to infection are dependent.

Let $M_{00}, M_{01}, M_{11}, M_{10}$ denote the likelihood factors corresponding to the four cases in Figure 5.5, i.e. the four factors in (5.4). Dropping the subject subscripts, let $F_I = F_I(C_i)$, $\bar{F}_I = 1 - F_I(C_i)$, and similarly for $F_S^\star$ and $\bar{F}_S^\star$. The structural assumptions allow us to replace the general bivariate quantiles arising from (5.5) by simple expressions that only use the univariate marginal distributions, which are

identifiable. Hence we get

$$
\begin{aligned}
M_{00} &= w\bar{F}_I\,\bar{F}_S^{\star} + (1-w)\bar{F}_I,\\
M_{01} &= w\bar{F}_I F_S^{\star},\\
M_{11} &= wF_I F_S^{\star} + (1-w)\int_0^C f_I(I)F_L(C-I)dI,\\
M_{10} &= wF_I\bar{F}_S^{\star} + (1-w)\int_0^C f_I(I)\bar{F}_L(C-I)dI.
\end{aligned}
$$

5.4.2 Bayesian Nonparametric Priors

The model is completed by introducing priors for the two unknown distributions, assuming nonparametric mixture models for both $f_I(I)$ and $f_S^{\star}(S)$,

$$
\begin{aligned}
f_I(I) &= \int \mathrm{N}(I \mid \boldsymbol{\theta}^{(I)})dH^{(I)}(\boldsymbol{\theta}^{(I)}) = \sum_{k=1}^{+\infty}\pi_k^{(I)}\mathrm{N}(I \mid \mu_k^{(I)}, \sigma_k^{(I)2})\\
f_S^{\star}(S) &= \int \mathrm{N}(I \mid \boldsymbol{\theta}^{(S)})dH^{(S)}(\boldsymbol{\theta}^{(S)}) = \sum_{k=1}^{+\infty}\pi_k^{(S)}\mathrm{N}(S \mid \mu_k^{(S)}, \sigma_k^{(S)2}),
\end{aligned}
\tag{5.8}
$$

where $\boldsymbol{\theta}^{(I)} = (\mu^{(I)}, \sigma^{(I)2})$ and $\boldsymbol{\theta}^{(S)} = (\mu^{(S)}, \sigma^{(S)2})$. Here $H^{(I)}(\cdot) = \sum_k \pi_k^{(I)}\delta_{\boldsymbol{\theta}^{(I)}}$, and similarly $H^{(S)}$, are the random mixing measures. The model is completed with a prior probability model on $H^{(I)}$ and $H^{(S)}$. Prior distributions on random probability measures are known as nonparametric Bayes (BNP) models.

Using a nonparametric prior on $H^{(I)}$ and $H^{(S)}$ the model becomes a mixture of normals with respect to the chosen random mixing measure. For example, in our implementation we assume a DP prior again, as in (5.3), now using two instances for f_I and $f_S^{\star}$. Alternatively, any other nonparametric Bayesian prior (e.g. James *et al.*, 2009) could be used. The following result gives the marginal distributions implied

by our construction.

Theorem 4. *The marginal distributions implied by model* (5.6) *with priors* (5.8) *are*

$$f_I(I) = \sum_{k=1}^{+\infty} \pi_k^{(I)} \mathrm{N}(I \mid \mu_k^{(I)}, \sigma_k^{(I)2}) \tag{5.9}$$

$$f_S(S) = w \sum_{k=1}^{+\infty} \pi_k^{(S)} \mathrm{N}(S \mid \mu_k^{(S)}, \sigma_k^{(S)2}) + (1-w) \sum_{k=1}^{+\infty} \pi_k^{(I)} \mathrm{EMG}(S \mid \mu_k^{(I)}, \sigma_k^{(I)2}, \lambda_L), \tag{5.10}$$

where $\mathrm{EMG}(\mu, \sigma^2, \lambda)$ *denotes the exponentially modified Gaussian distribution first introduced in Grushka (1972).*

Proof. *We begin by calculating the marginal distribution for the infection times as*

$$\begin{aligned} f_I(I) &= \int f_{I,S}(I,S) dS \\ &= w f_I(I) + (1-w) f_I(I) \int f_L(S-I) dS \\ &= w f_I(I) + (1-w) f_I(I) \int_I^{+\infty} \lambda_L e^{-\lambda(S-I)} dS = f_I(I) \\ &= \sum_{k=1}^{+\infty} \pi_k^{(I)} \mathrm{N}(I \mid \mu_k^{(I)}, \sigma_k^{(I)2}). \end{aligned}$$

The marginal distribution for the symptoms times is

$$\begin{aligned} f_S(S) &= \int f_{I,S}(I,S) dI \\ &= w f_S(S) + (1-w) \int_{-\infty}^{S} f_I(I) f_L(S-I) dI \\ &= w \sum_{k=1}^{+\infty} \pi_k^{(S)} \mathrm{N}(S \mid \mu_k^{(S)}, \sigma_k^{(S)2}) + \\ &\quad (1-w) \sum_{k=1}^{+\infty} \pi_k^{(I)} \lambda_L \exp\left\{ \frac{\lambda_L}{2} (\lambda_L \sigma_k^{2(I)} + 2\mu_k^{(I)} - 2S) \right\} \Phi\left(\frac{S - \mu_k^{(I)} - \lambda_L \sigma_k^{(I)2}}{\sigma_k^{(I)}} \right). \end{aligned}$$

■

Model (5.6) together with (5.8) and (5.3) for $p(C_i \mid S_i)$ defines the proposed **bivariate BNP-CS model** for current status data.

One of the reasons for the wide use of BNP mixtures like (5.8) is the induced prior on a random partition. Consider $I_i \sim f_I$, $i = 1, \ldots, n$. Under model (5.8) we can introduce latent indicators, say $r_i^{(I)}$, and write instead

$$p(I_i \mid r_i^{(I)} = k) = \mathrm{N}(\mu_k^{(I)}, \sigma_k^{(I)2}) \quad \text{and} \quad p(r_i^{(I)} = k) = \pi_k^{(I)}.$$

The $r_i^{(I)}$'s can be interpreted as cluster membership indicators. We see then how this formulation implicitly defines a probability model $p(\mathbf{r}^{(I)})$ on a partition $\mathbf{r}^{(I)} = (r_1^{(I)}, \ldots, r_n^{(I)})$. Two observations are clustered together if they are assigned the same group-specific parameters $\boldsymbol{\theta}_k = (\mu_k, \sigma_k^2)$, where for brevity we now omit the superscript (I). Recall the indicators v_i in (5.7). Without loss of generality assume that $v_i = 1$ (symptoms due to other causes) for $i = 1, \ldots, n_1$, and $v_i = 0$ (symptoms due to disease), $i = n_1 + 1, \ldots, n$. Similar to $p(\mathbf{r}^{(I)})$ we get a random partition $p(\mathbf{r}_\star^{(S)})$ induced by sampling from $f_S^\star(\cdot)$ for patients $i = 1, \ldots, n_1$. For $i = n_1 + 1, \ldots, n$ we have $S_i = I_i + L_i$ with the infection times I_i subject to the already described partition $\mathbf{r}^{(I)}$, and L_i i.i.d. under the assumed parametric model for the lag times $L_i = S_i - I_i$. In words, under the proposed model, the clustering structures $\mathbf{r}^{(S)}$ and $\mathbf{r}_\star^{(S)}$ for symptoms due to infection and for symptoms due to other causes, respectively, are modeled separately and are independent. In fact, symptoms due to infection inherit the clustering structure $\mathbf{r}^{(I)}$, which is induced by the marginal distribution for the infection times.

In order to cluster grouped data, other approaches have been proposed (Teh *et al.*, 2005; Rodriguez *et al.*, 2008; Camerlenghi *et al.*, 2019; Argiento *et al.*, 2020). These strategies allow for the possibility of sharing atoms of the random probability measures across groups, thus borrowing information and yielding more precise inference. However, the random partition is not the main inference target here and we shall therefore not further explore such alternatives.

5.4.3 Regression on Covariates

We now add covariate effects in the proposed nonparametric model. In the context of model (5.8) this takes the form of replacing $H^{(I)}$ and $H^{(S)}$ by families of random probability measures (r.p.m). That is, we introduce a family $\{H^{(I)}_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\}$, and similarly for $H^{(S)}$. Here $\mathbf{x}$ are patient specific covariates, and we replace $H^{(I)}$ and $H^{(S)}$ by $H^{(I)}_{\mathbf{x}_i}$ and $H^{(S)}_{\mathbf{x}_i}$ for patient i in equation (5.8). Dropping for the moment the superscript for easier exposition, let $\mathcal{H} = \{H_{\mathbf{x}} = \sum_k \pi_{xk}\delta_{\mu_{xk}}, \mathbf{x} \in \mathcal{X}\}$ denote a family of r.p.m.'s indexed by $\mathbf{x}$. The most widely used class of priors on families like $\mathcal{H}$ are dependent DP (DDP) models (MacEachern, 1999). The DDP construction implies marginally for each $H_{\mathbf{x}}$ a DP prior, and allows for the desired dependence across $\mathbf{x}$. The definition of the marginal DP implies that the μ_{xk}'s are independent across k and that the weights have stick-breaking priors, but it does not restrict the distribution across $\mathbf{x}$. This is what the DDP construction exploits to borrow information across covariate values. The DDP induces dependence across $\mathbf{x}$ through the atoms μ_{xk} and/or the weights π_{xk} of the marginal r.p.m.'s.

In the Partner Notification study the predictors are $\mathbf{x}_i = \{\text{gender}, \text{arm}, \text{age}\} \in$

$\{0;1\}^2 \times \mathbb{R}^+$, i.e. two binary and one continuous covariate. We use a simple ANOVA structure to induce dependence of μ_{xk} across $\mathbf{x}$ and common weights π_h. DDP models with ANOVA-type dependence across categorical factors are introduced as the ANOVA-DDP in De Iorio *et al.* (2004) and then extended to continuous covariates in De Iorio *et al.* (2009). The dependence structure of the random probability measures $H_{\mathbf{x}}$ is modeled by constructing the atoms as $\mu_{xk} = \delta_k + \alpha_k x_1 + \beta_k x_2 + \gamma_k x_3$. The interpretation of the linear model coefficients $\mathbf{m}_k = (\delta_k, \alpha_k, \beta_k, \gamma_k)^\intercal$ is exactly as in an ANOVA model, inducing the desired dependence of $H_{\mathbf{x}}$ across $\mathbf{x}$ by sharing, for example, the same β_k for any two covariate vectors $\mathbf{x}$ and $\mathbf{x}'$ that share the same x_2. Finally, using a design vector $\mathbf{d}_i = (1, x_{i1}, x_{i2}, x_{i3})^\intercal$ to select the desired ANOVA effects we can write $\mu_{x_i k} = \mathbf{d}_i^\intercal \mathbf{m}_k$ to get $H_{\mathbf{x}_i} = \sum_{k=1}^{+\infty} \pi_k \delta_{\mathbf{d}^\intercal \mathbf{m}_k}$. Defining $\boldsymbol{\theta}_k = (\mathbf{m}_k, \sigma_k^2)^\intercal$ to allow for a mixture also with respect to the kernel variances, we can define one common mixing measure

$$H(\cdot) = \sum_{k=1}^{+\infty} \pi_k \delta_{\boldsymbol{\theta}_k},$$

and push the linear model into the mixture kernel in (5.8). We now add back the superscripts (I) and (S) for the two models $H^{(I)}$ and $H^{(S)}$ in equation (5.8). Using $H^{(I)}$ the marginal distribution $f_I(I_i \mid \mathbf{x}_i)$ can thus be rewritten equivalently as a DP mixture of linear models, now using a single mixing measure H for all $\mathbf{x}$ (linear dependent DDP, Jara *et al.*, 2010)

$$f_I(I_i \mid \mathbf{x}_i) = \int \mathrm{N}(I_i \mid \mathbf{d}_i^\intercal \mathbf{m}^{(I)}, \sigma^{(I)2}) dH^{(I)}(\boldsymbol{\theta}^{(I)}) \quad \text{with} \quad H^{(I)} \sim \mathrm{DP}(M^{(I)}, H_0^{(I)}). \tag{5.11}$$

Another instance of the same model is used for the marginal distribution of symptoms

due to other causes $f_S^\star(S_i \mid \mathbf{x}_i)$. The full model is

$$C_i \mid S_i, \lambda = \min\{S_i + \text{Exp}(\lambda); \text{Unif}(A, B)\}$$

$$(S_i, I_i) \mid \boldsymbol{\theta}^{(S)}, \boldsymbol{\theta}^{(I)}, w, \lambda_L \sim f_{IS}(I, S).$$

using (5.11) for f_I and similarly for $f_S^\star$. The complete model now defines a **bivariate BNP-CS survival regression**. Using the stick-breaking representation, the DP priors on $H^{(I)}$ and $H^{(S)}$ can be written as follows. Using superscripts $E \in \{I, S\}$ to refer to the construction of f_I and $f_S^\star$ respectively, we have

$$\{\mathbf{m}_k^{(E)}, \sigma_k^{(E)2}\}_{k=1}^{+\infty} \overset{\text{iid}}{\sim} H_0^{(E)} = \text{N}(\mathbf{m}_k^{(E)} \mid \mathbf{m}_0^{(E)}, \Sigma_0^{(E)}) \times \text{IG}(\sigma_k^{(E)2} \mid a_\sigma^{(E)}, b_\sigma^{(E)})$$

$$\boldsymbol{\pi}^{(E)} \mid M^{(E)} \sim \text{SB}(M^{(E)}); \quad M^{(E)} \sim \text{Ga}(a_M, b_M),$$

and $\lambda \sim \text{Ga}(a_\lambda, b_\lambda)$, $\lambda_L \sim \text{Ga}(a_L, b_L)$, $w \sim \text{Beta}(a_w, b_w)$. This completes the model construction. Hyperparameter choices are described in Section 5.5.2.

For later reference we note that the random probability measures $H^{(I)}(\boldsymbol{\theta}^{(I)})$ and $H^{(S)}(\boldsymbol{\theta}^{(S)})$ that serve as the mixing measure in (5.11) are multivariate distributions for $\boldsymbol{\theta}^{(I)} = (\mathbf{m}^{(I)}, \sigma^{(I)2})^\intercal = (\delta^{(I)}, \alpha^{(I)}, \beta^{(I)}, \gamma^{(I)}, \sigma^{(I)2})^\intercal$, and similarly for $\boldsymbol{\theta}^{(S)}$. Let

$$H_\beta^{(I)} = \sum_{k=1}^{+\infty} \pi_k^{(I)} \delta_{\beta_k^{(I)}} \tag{5.12}$$

denote the implied univariate marginal for the ANOVA effect $\beta^{(I)}$. Analogous notation can be used for $H_\beta^{(S)}$ and any of the other ANOVA effects. We will later use inference on $H_\beta^{(E)}$, $E \in \{I, S\}$, to summarize inference on the treatment effect.

5.5 Posterior Inference

To implement posterior inference under a Dirichlet process mixture model, the two main strategies are marginal (Escobar and West, 1995; MacEachern and

Müller, 1998; Neal, 2000) and conditional (Ishwaran and James, 2001; Kalli *et al.*, 2011) MCMC posterior simulation. In our implementation, we employ the latter. In particular, we rewrite the mixture model as a hierarchy by explicitly introducing the latent cluster membership variables $\mathbf{v}$, $\mathbf{r}^{(I)}$ and $\mathbf{r}_{\star}^{(S)}$. Moreover, we impute the latent symptoms and infection times from their corresponding full conditionals. We use efficient sampling for truncated normal distributions, originally proposed in Geweke (1991a). This allows us to use standard algorithms for inference under a DPM.

The total masses $M^{(I)}$ and $M^{(S)}$ for the two random probability measures are included in the MCMC scheme and assigned Gamma priors, as recommended in Escobar and West (1995). Moreover, we put priors on the hyperparameters for the base measures $H_0^{(I)}$ and $H_0^{(S)}$. Additional details of the algorithm are described below.

5.5.1 MCMC Algorithm

Posterior inference for the bivariate survival regression model, described in Section 5.4, is based on a posterior Monte Carlo sample generated using a Gibbs sampler simulation. In what follows, $\boldsymbol{\zeta}$ denotes a generic variable that collects all other variables not explicitly mentioned, including the data.

The algorithm imputes the latent times to symptoms and times to infection. Due to space constraint in the table in Algorithm 3, we detail here how these parameters can be sampled. The times until symptoms due to the infection are sampled

from

$$p(S_i \mid \boldsymbol{\zeta}) \propto \text{Exp}(\lambda_L - \lambda)\big|_{I_i}^{C_i} \qquad \text{if } \Delta_{S_i} = 1, v_i = 0$$

$$p(S_i \mid \boldsymbol{\zeta}) \propto \text{Exp}(\lambda_L)\big|_{\max\{C_i, I_i\}}^{+\infty} \quad \text{if } \Delta_{S_i} = 0, v_i = 0.$$

Times until symptoms due to other causes are sampled from

$$p(S_i \mid r_{i\star}^{(S)} = k, \boldsymbol{\zeta}) \propto \text{N}(\mu_k^{(S)}, \sigma_k^{(S)2})\big|_{-\infty}^{C_i} \quad \text{if } \Delta_{S_i} = 1, v_i = 1$$

$$p(S_i \mid r_{i\star}^{(S)} = k, \boldsymbol{\zeta}) \propto \text{N}(\mu_k^{(S)}, \sigma_k^{(S)2})\big|_{C_i}^{+\infty} \quad \text{if } \Delta_{S_i} = 0, v_i = 1.$$

Times until infection are sampled from

$$p(I_i \mid r_i^{(I)} = k, \boldsymbol{\zeta}) \propto \text{N}(\mu_k^{(I)} + \lambda\sigma_k^{(I)2}, \sigma_k^{(I)2})\big|_{-\infty}^{\min\{C_i, S_i\}} \quad \text{if } \Delta_{I_i} = 1, v_i = 0$$

$$p(I_i \mid r_i^{(I)} = k, \boldsymbol{\zeta}) \propto \text{N}(\mu_k^{(I)} + \lambda\sigma_k^{(I)2}, \sigma_k^{(I)2})\big|_{C_i}^{S_i} \qquad \text{if } \Delta_{I_i} = 0, v_i = 0$$

$$p(I_i \mid r_i^{(I)} = k, \boldsymbol{\zeta}) \propto \text{N}(\mu_k^{(I)}, \sigma_k^{(I)2})\big|_{-\infty}^{C_i} \qquad \text{if } \Delta_{I_i} = 1, v_i = 1$$

$$p(I_i \mid r_i^{(I)} = k, \boldsymbol{\zeta}) \propto \text{N}(\mu_k^{(I)}, \sigma_k^{(I)2})\big|_{C_i}^{+\infty} \qquad \text{if } \Delta_{I_i} = 0, v_i = 1.$$

As mentioned above, we use a truncated approximation to the infinite mixture model. Let K_{max} be the truncation level (in the following, we fix $K_{max} = 40$). We describe here the case without covariates, although the regression terms are straightforward to include in the algorithm. We also do not include the update for the base measure hyperparameters as it consists of a simple normal full conditional. The sampler for the proposed model of Section 5.4 comprises the steps outlined in Algorithm 3.

5.5.2 Prior Hyper-parameters and MCMC Initializations

The parameters λ and λ_L were assigned Gamma priors $\lambda \sim \text{Ga}(a_\lambda, b_\lambda)$, $\lambda_L \sim \text{Ga}(a_L, b_L)$. The hyperparameters were chosen to imply the 95% prior cred-

Algorithm 3 (Gibbs Sampler)

Updating the symptoms parameters

1: For $i = 1, \ldots, n$, sample the latent times until symptoms S_i as described above.
2: For $k = 1, \ldots, K_{max}$, sample the group specific parameters $\mu_k^{(S)}, \sigma_k^{(S)2}$ as

$$\mu_k^{(S)} \mid \boldsymbol{\zeta} \sim p_0(\mu_k^{(S)}) \prod_{i \text{ s.t.} r_{i\star}^{(S)}=k} p(S_i \mid \mu_k^{(S)}, \sigma_k^{(S)2}),$$

$$\sigma_k^{(S)2} \mid \boldsymbol{\zeta} \sim p_0(\sigma_k^{(S)2}) \prod_{i \text{ s.t.} r_{i\star}^{(S)}=k} p(S_i \mid \mu_k^{(S)}, \sigma_k^{(S)2}).$$

3: For $i = 1, \ldots, n_1$, sample the cluster membership indicators $r_{i\star}^{(S)}$ as

$$p(r_{i\star}^{(S)} = k \mid \boldsymbol{\zeta}) \propto \pi_k^{(S)} \mathrm{N}(S_i \mid \mu_k^{(S)}, \sigma_k^{(S)2}).$$

4: For $k = 1, \ldots, K_{max}$, update the weights $\pi_k^{(S)} = q_k^{(S)} \prod_{\ell<k}(1 - q_\ell^{(S)})$, where $q_k \mid \boldsymbol{\zeta} \sim \mathrm{Beta}(1 + n_k^{(S)}, M^{(S)} - \sum_{\ell=k+1}^{K_{max}} n_\ell^{(S)})$ and $n_k^{(S)} = \sum_i \mathbb{1}\{r_{i\star}^{(S)} = k\}$.

Updating the infection parameters

5: For $i = 1, \ldots, n$, sample the latent infection times I_i as described above.
6: For $k = 1, \ldots, K_{max}$, sample the group specific parameters $\mu_k^{(I)}, \sigma_k^{(I)2}$ as

$$\mu_k^{(I)} \mid \boldsymbol{\zeta} \sim p_0(\mu_k^{(I)}) \prod_{i \text{ s.t.} r_i^{(I)}=k} p(I_i \mid \mu_k^{(I)}, \sigma_k^{(I)2}),$$

$$\sigma_k^{(I)2} \mid \boldsymbol{\zeta} \sim p_0(\sigma_k^{(I)2}) \prod_{i \text{ s.t.} r_i^{(I)}=k} p(I_i \mid \mu_k^{(I)}, \sigma_k^{(I)2}).$$

7: For $i = 1, \ldots, n$, sample the cluster membership indicators $r_i^{(I)}$ as

$$p(r_i^{(I)} = k \mid \boldsymbol{\zeta}) \propto \pi_k^{(I)} \mathrm{N}(I_i \mid \mu_k^{(I)}, \sigma_k^{(I)2}).$$

8: For $k = 1, \ldots, K_{max}$, update the weights $\pi_k^{(I)} = q_k^{(I)} \prod_{\ell<k}(1 - q_\ell^{(I)})$, where $q_k \mid \boldsymbol{\zeta} \sim \mathrm{Beta}(1 + n_k^{(I)}, M^{(I)} - \sum_{\ell=k+1}^{K_{max}} n_\ell^{(I)})$ and $n_k^{(I)} = \sum_i \mathbb{1}\{r_i^{(I)} = k\}$.

Updating the global parameters

9: Update the dependent censoring parameter λ with a M-H transition probability using the target distribution

$$p(\lambda \mid \boldsymbol{\zeta}) \propto p_0(\lambda) \prod_{i \text{ s.t.} \Delta_{S_i}=1} p(C_i \mid S_i, \lambda).$$

10: Sample the indicator for dependent symptoms $v_i \mid \boldsymbol{\zeta} \sim \text{Be}(p_i)$, where $p_i = p^\star_{i,1}/(p^\star_{i,0} + p^\star_{i,1})$, $p^\star_{i,1} = w \sum_{k=1}^{+\infty} \pi_k^{(S)} \text{N}(S_i \mid \mu_k^{(S)}, \sigma_k^{(S)2})$, $p^\star_{i,0} = (1-w)\lambda e^{-\lambda(S_i - I_i)}$.

11: Sample the probability for dependent symptoms $w \mid \boldsymbol{\zeta} \sim \text{Beta}(a_w + \sum_i v_i, b_w + n - \sum_i v_i)$.

12: Sample the latency parameter $\lambda_L \mid \boldsymbol{\zeta} \sim \text{Ga}\{a_L + n - \sum_i v_i, b_L + \sum_{i \text{ s.t.} v_i=0} (S_i - I_i)\}$.

ible intervals for the latency times to be $[0.05, 9]$ days, yielding $a_\lambda = a_L = 10$, $b_\lambda = b_L = 20$. The proportion w of individuals with symptoms due to other causes has a $\text{Beta}(a_w, b_w)$ prior. The hyperparameters were chosen so that $a_w = b_w = 1$, i.e. a uniform prior.

The total masses for the two random probability measures $M^{(I)}$ and $M^{(S)}$ are included in the MCMC scheme and assigned Gamma priors, as recommended in Escobar and West (1995). We use $a_M = 10, b_M = 1$ for both of them. Recall the base measures $H_0^{(E)}$, $E \in \{I, S\}$,

$$\{\mathbf{m}_k^{(E)}, \sigma_k^{(E)2}\}_{k=1}^{+\infty} \overset{\text{iid}}{\sim} H_0^{(E)} = \text{N}(\mathbf{m}_k^{(E)} \mid \mathbf{m}_0^{(E)}, \Sigma_0^{(E)}) \times \text{IG}(\sigma_k^{(E)2} \mid a_\sigma^{(E)}, b_\sigma^{(E)}).$$

We use vague priors for $m_{0j} \overset{\text{iid}}{\sim} \text{Normal}(0, 100^2)$, $\Sigma_0^{(E)} = \text{diag}(\sigma_{0j}^2)_{j=1}^p$, and $\sigma_{0j}^2 \overset{\text{iid}}{\sim} \text{Inv-Ga}(1,1)$, $b_\sigma^{(E)} \sim \text{Ga}(1,1)$, whereas we fix $a_\sigma^{(E)} = 1$.

The initialization for the partitions of infection times and times until symptoms were obtained by using a K-means algorithm on the censoring times, with

$K = 5$. The group-specific location and scale parameters were initialized to the corresponding maximum likelihood estimators. The remaining parameters were initialized from their priors.

The algorithm proves to be very robust to both the prior specification and to the initialization.

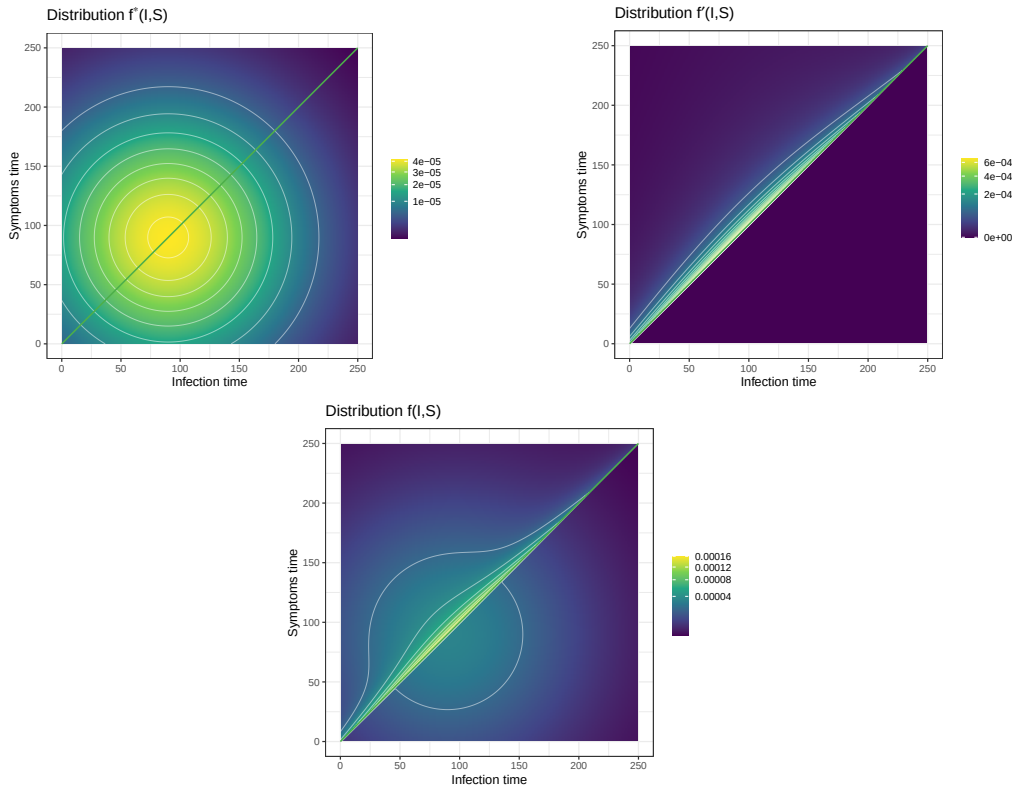


Figure 5.6: Prior mean density estimate for $f^{\star}_{IS}$, f'_{IS} and f_{IS} corresponding to the baseline covariate levels (male, control group, mean age). The green line corresponds to the 45° line, i.e. $I = S$.

5.5.3 Convergence Diagnostics

This section presents some MCMC convergence diagnostics for the proposed Gibbs sampler. The results presented here are obtained on the real data analysis.

The Geweke test (Geweke, 1991b) for stationarity of the chains, which formally compares the means of an early vs a later part of a Markov chain (by default the first 10% and the last 50%), is also performed. If the samples were from the stationary distribution of the chain, the two means are equal and Geweke's statistic has an asymptotically standard normal distribution. We perform the Geweke test to assess convergence using the global parameters, i.e. those that are not affected by label switching. Both the exponential parameters λ and λ_L as well as the proportion of patients with symptoms due to other causes w, have very stable traceplots (see Figure 5.7) and fail to reject the null hypothesis of stationarity of the corresponding chains.

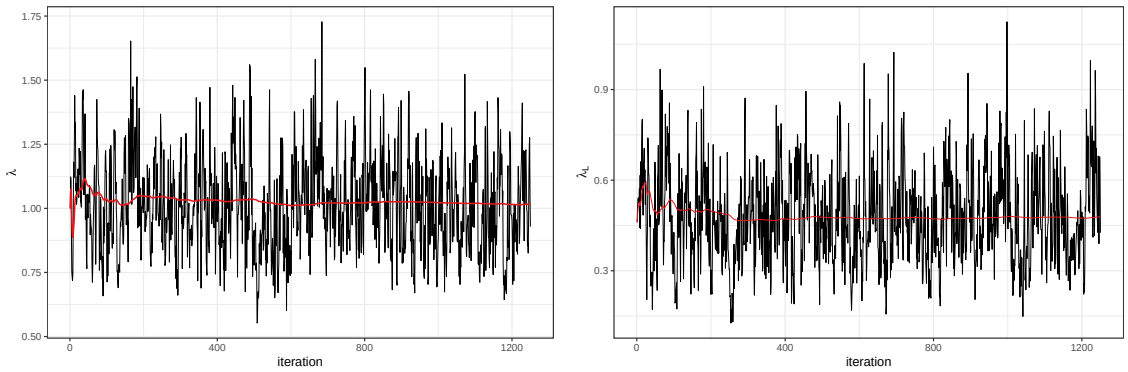


Figure 5.7: Trace plots of the exponential parameters λ and λ_L. In each panel, the solid red line shows the running mean.

5.5.4 Software, Runtime, etc.

We programmed in `R` interfaced with `C++`. A total of 35000 MCMC iterations were run with the initial 10000 iterations discarded as burn-in. The chain was subsequently thinned every 20 iterations. The MCMC algorithm takes 10 minutes on a Macbook laptop with 8 Gb RAM.

5.6 Simulation Studies

In this section, we discuss the results of some simulated numerical experiments. In designing the simulation scenarios, we have tried to closely mimic our recurrent infection dataset. We thus chose $n \in \{250, 1000\}$ participants being followed in the time window $[A, B] = [0, 200]$. We also simulate two covariates: a *binary* X_1 and a *continuous* X_2. The underlying distribution for the infection times is a mixture of two linear models with weights $\boldsymbol{\pi}^{(I)} = (0.6, 0.4)^\intercal$, location parameters $\mathbf{m}^{(I)} = [\mathbf{m}_1^{(I)} \quad \mathbf{m}_2^{(I)}] = ((40, -5, 0)^\intercal, (100, -10, -15)^\intercal)$ and scale parameters

$\boldsymbol{\sigma}^{(I)2} = (10^2, 10^2)^\intercal$. The distribution for the symptom times due to other causes is a mixture of two linear models with weights $\boldsymbol{\pi}^{(S)} = (0.4, 0.6)^\intercal$, location parameters $\mathbf{m}^{(S)} = [\mathbf{m}_1^{(S)} \quad \mathbf{m}_2^{(S)}] = ((70, 0, 20)^\intercal, (110, -5, 0)^\intercal)$ and scale parameters $\boldsymbol{\sigma}^{(S)2} = (10^2, 20^2)^\intercal$. Figure 5.8 shows the results when the proportion of patients with symptoms due to other causes is $w = 0.75$. In this simulation, the latency time parameter between infection times and symptom times λ_L as well as the dependent censoring parameter λ are chosen to be 0.2. As one can see in Figure 5.8 the underlying true bivariate density is recovered well by our method. Estimates for other relevant parameters are not reported here, but were also very accurate.

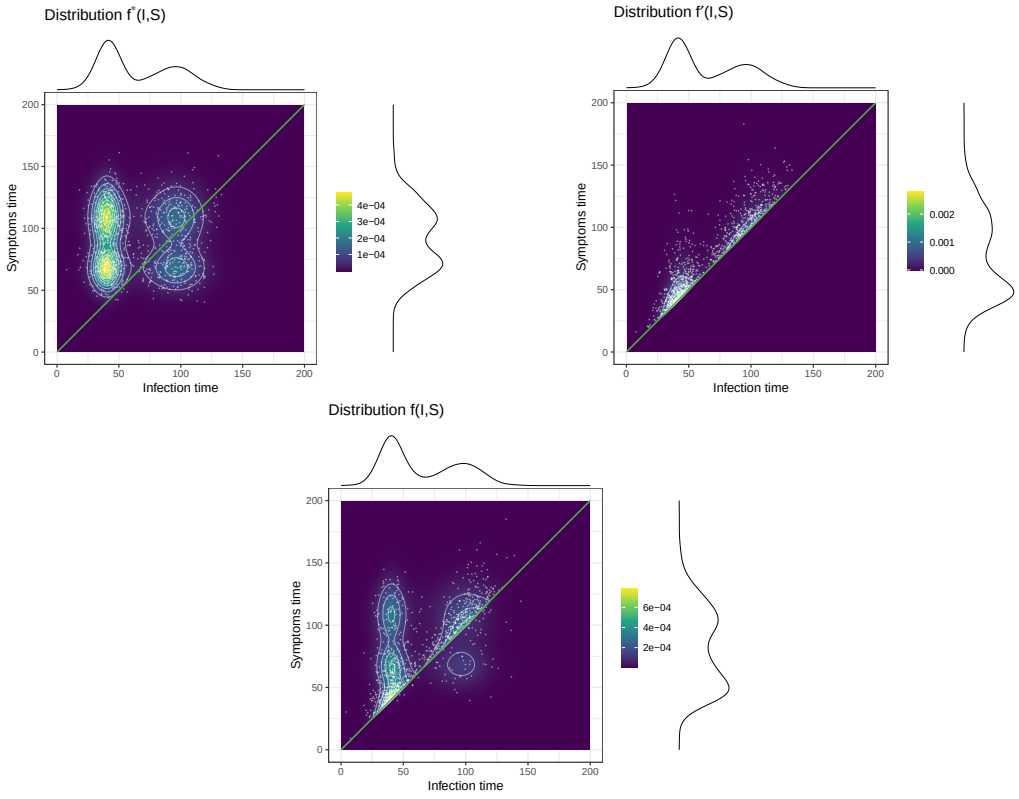


Figure 5.8: Results for simulated data: Posterior mean density estimate for $f^{\star}_{IS}$, f'_{IS} and f_{IS} corresponding to the baseline covariate levels. The green line is the 45° line $I = S$. The corresponding marginal distributions are shown on the top and right side of the density plot. The white points are a sample of the true latent times corresponding to the same covariate levels.

We show how the proposed model compares with two independent ANOVA-DDP models for the marginal distributions, as well as with the bivariate Gumbel model in a variety of scenarios. In particular, we design three studies when the simulated data have the following features: (I) independent censoring ($\lambda \approx 0$) and dependent symptoms ($w = 0.5$), (II) dependent censoring ($\lambda = 0.2$) and indepen-

dent symptoms ($w = 1$), and (III) dependent censoring ($\lambda = 0.2$) and dependent symptoms ($w = 0.5$). All the other parameters are kept fixed as described above.

To evaluate model performance, we measure how well the models are able to recover the functional form of the survival curves for the two marginal distributions. In particular, we use the mean integrated squared error (MISE). The MISE for estimating $f(t)$ by $\hat{f}(t)$ is defined as

$$\text{MISE} = \mathbb{E}\left[\int \left\{f(t) - \hat{f}(t)\right\}^2 dt\right]$$

We estimate the MISE by averaging the estimated integral across D simulated data sets as $\text{MISE}_{\text{est}} = \frac{1}{D}\sum_{d=1}^{D}\sum_{i=1}^{N}\Delta_i\{f(t_i) - \hat{f}^{(d)}(t_i)\}^2$, where $\Delta_i = t_i - t_{i-1}$, $\{t_i\}_{i=1}^{N}$ are a set of grid points on the range of the data and $\hat{f}^{(d)}$ is the estimated function of interest for data set d. In Table 5.1, the reported estimated MISEs are based on $D = 50$ simulated data sets. This simulation shows how the proposed model outperforms the marginal ANOVA-DDP models and the bivariate Gumbel model by exploiting the dependence structure of the data under a wide variety of data generating mechanisms.

5.7 Partner Notification Study - Results

We apply the proposed model for inference in the Partner Notification study described in Section 5.2. The primary inference goal is to understand the effect of covariates, in particular treatment assignment, on the joint distribution of the two latent times of interest. Furthermore, we are interested in assessing what factors drive time to reinfection and how time to symptoms onset of these cases can improve

	Sample Size	Distr.	De Iorio et al.	Bivariate Gumbel	Our method
(I)	$n = 250$	Inf.	1.64 (0.92, 3.01)	4.01 (3.14, 5.59)	1.10 (0.09, 2.24)
		Sym.	2.98 (1.11, 5.01)	6.15 (5.31, 8.77)	1.33 (0.18, 3.72)
	$n = 1000$	Inf.	1.32 (0.73, 1.90)	3.76 (3.19, 4.54)	0.50 (0.04, 1.80)
		Sym.	2.32 (1.19, 3.25)	5.99 (5.31, 6.99)	1.30 (0.54, 2.66)
(II)	$n = 250$	Inf.	0.96 (0.74, 1.56)	3.44 (3.08, 4.59)	0.99 (0.13, 2.07)
		Sym.	8.44 (5.21, 12.30)	11.75 (9.18, 18.01)	**0.76** (0.22, 2.16)
	$n = 1000$	Inf.	0.80 (0.50, 1.10)	3.12 (3.03, 3.41)	**0.19** (0.05, 0.50)
		Sym.	8.18 (6.28, 10.32)	10.74 (9.58, 12.49)	**0.12** (0.02, 0.37)
(III)	$n = 250$	Inf.	4.45 (3.00, 6.30)	4.24 (3.09, 5.79)	**0.45** (0.08, 1.14)
		Sym.	9.82 (6.70, 13.20)	8.08 (5.72, 12.15)	**0.24** (0.03, 0.81)
	$n = 1000$	Inf.	4.10 (3.18, 4.96)	3.96 (3.24, 4.81)	**0.13** (0.01, 0.35)
		Sym.	9.94 (8.44, 11.71)	7.98 (6.31, 10.06)	**0.05** (0.01, 0.15)

Table 5.1: Results for simulated data: Estimated median integrated squared error (MISE_{est}) performance of the survival regression model described in Section 5.4 compared with the method of De Iorio et al. (2009) and with a bivariate Gumbel survival regression. We have reported here the MISE values for estimating the two marginal distributions (infection and symptoms, respectively) corresponding to the baseline covariate levels. In parenthesis, the 95% credible intervals for the MISE values are reported. When a method significantly outperforms the other, the corresponding MISE value is highlighted in bold.

such estimation.

Inference under the proposed model includes the full joint distribution of latent times to symptoms and infection times. Figure 5.9 shows the posterior estimated distribution $f_{IS}(I, S)$ and the two components $f^{\star}_{IS}(I, S)$, $f'_{IS}(I, S)$ corresponding to a 'baseline' covariate combination (male, control group, median age). There is significant probability mass in the lower triangle ($S < I$) that is not concentrated around the 45° line but is quite spread out. Instead, for the constrained component ($S > I$)

the probability mass is concentrated very close to the 45° line. In other words, most of the inferred symptoms times due to infection concentrate in $I < S < I + 10$. This is simply reflecting that symptoms due to the infection follow shortly after the disease onset.

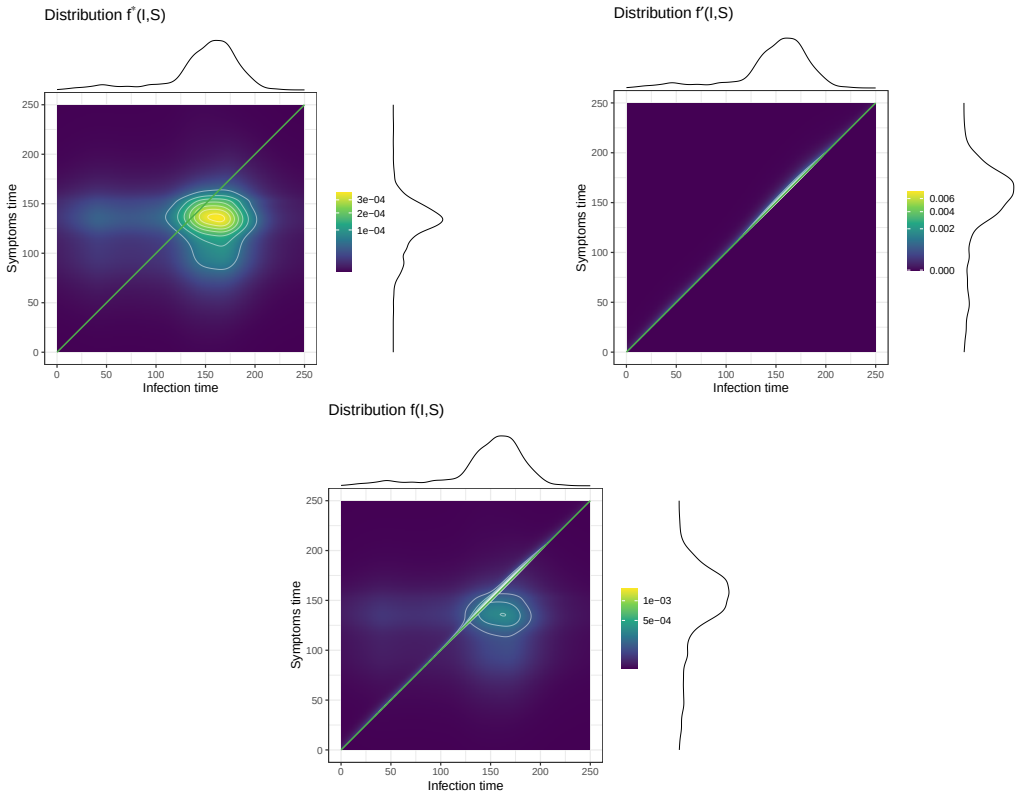


Figure 5.9: Results: Posterior mean density estimate for $f^{\star}_{IS}$, f'_{IS} and f_{IS} corresponding to the baseline covariate levels (male, control group, mean age). The green line corresponds to the 45° line, i.e. $I = S$. The corresponding marginal distributions are shown on the top and right side of the density plot.

These results differ from the prior expectation implied by our choice of the hyperparameters. As one can see in Figure 5.6, the prior is much more diffuse than the posterior density estimate.

To show the estimated covariate effects, we could compare estimated survival functions for different combinations of the predictors. Alternatively, we can report posterior estimated marginal distributions of the ANOVA effects, for example $H_\beta^{(E)}$, $E \in \{I, S\}$ from (5.12). These are the univariate marginal distributions of the treatment effect in the DDP model, and concisely summarize the change of the bivariate survival distribution with respect to treatment versus control. The top center panel in Figure 5.10 shows the posterior estimated distributions $\mathbb{E}(H_\beta^{(I)} \mid \text{data})$, and similarly for other regression effects. Two significant effects can be detected. Importantly, treatment delays reinfection times, confirming what was found in an earlier analysis in Sal y Rosas and Hughes (2011). Moreover, gender has an effect on the time to symptoms due to other causes, with women seeking hospital visits earlier, when the visit is prompted by symptoms. This might be simply due to the fact that women are more aware of their symptoms and are more inclined to hospital visits, suggesting that a health education campaign for men might improve their health outcome. Age has also been found to have a weak effect: younger individuals have shorter times to reinfection, possibly due to their more risky behaviour.

Two parameters of the model, namely λ_L and λ, can give insights into how long it takes for participants to develop symptoms and to seek a visit to the hospital. In particular, the 95% credible interval for the exponential parameter λ is $[0.70, 1.42]$, suggesting that people seek a doctor visit, on average, one day after onset of symptoms. Moreover, the 95% credible interval for the exponential parameter λ_L is $[0.22, 0.80]$, which implies that patients develop symptoms due to infection, on average, 2.5 days after reinfection. Inference includes an estimate for the proportion

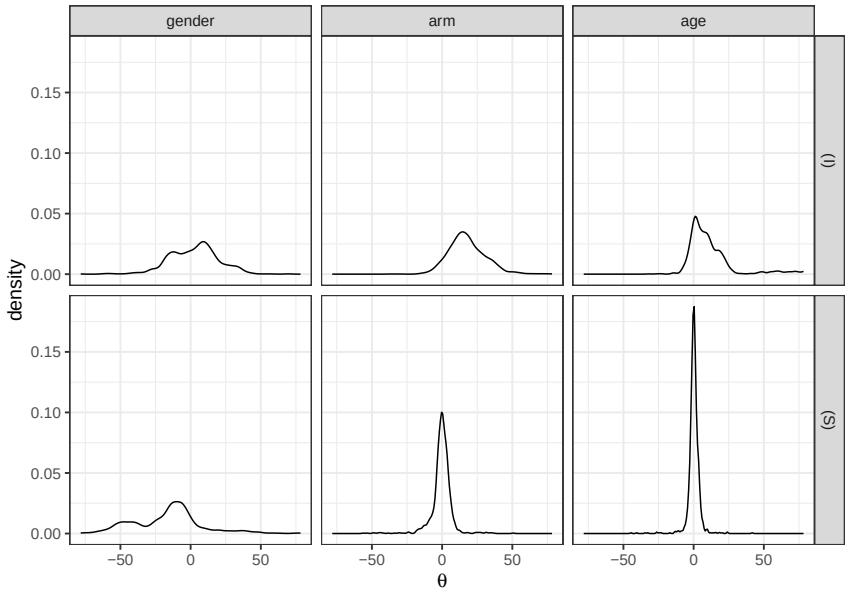


Figure 5.10: Results: Estimated distributions H_α, H_β and H_γ for the regression coefficients α (left), β (middle) and γ (right) under f_I (top panels) and $f_S^\star$ (bottom panels).

of patients that experience symptoms due to the infection, in our notation $1-w$. The posterior mean of such proportion is 0.18 (95% CI: [0.12, 0.24%]). This is coherent with what we see empirically in the data. There are more observed symptoms than observed infections, which implies that most of the symptoms should be attributed to other causes. This finding has important practical implications as it can help better planning for the treatment of patients.

By way of comparison we carry out alternative inference, first using a model with flexible marginal distributions that assumes independence between the two events, and then using a model with parametric marginal distributions that instead

implies dependence between the events. For the first strategy, we implement inference under two independent linear dependent Dirichlet process (LDDP) mixture of survival models for the marginal distributions of infection and symptoms times. This method is described in De Iorio *et al.* (2009) and implemented in the `DPpackage` (Jara *et al.*, 2011). For a fair comparison, we used the same prior specifications for the shared parameters under the two models. The results are shown in Figure 5.11. Note the inappropriate posterior shrinkage of probability mass toward the extremes.

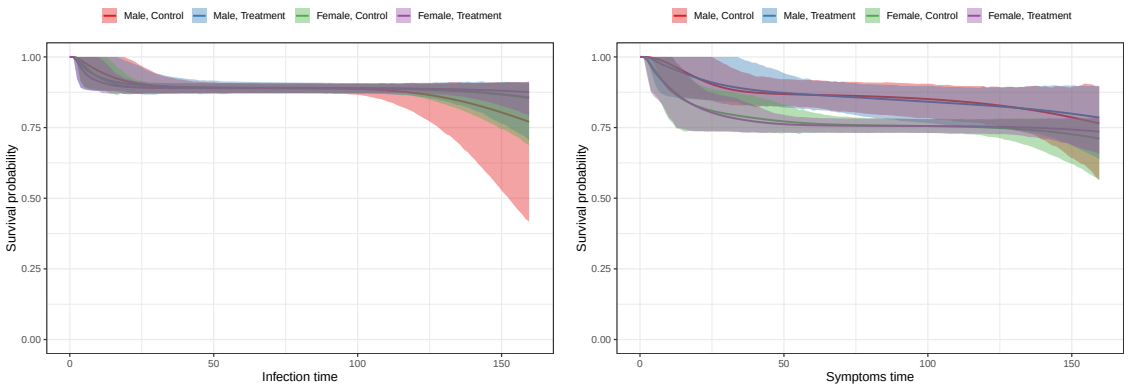


Figure 5.11: Results: LDDP estimated survival curves for infection times (left panel) and times until symptoms (right panel) corresponding to the possible combinations of the binary covariates *gender* and *treatment* fixing the predictor *age* to the average age in the sample.

The second comparison uses a bivariate Gumbel model (Gumbel, 1960). Two variables (I, S) have a Gumbel bivariate exponential distribution if their probability density function is

$$f(I,S) = \lambda_I e^{-\lambda_I I} \lambda_S e^{-\lambda_S S}[1 + \alpha\{1 - 2e^{-\lambda_I I}\}\{1 - 2e^{-\lambda_S S}\}],$$

where $-1 \leq \alpha \leq 1$ is a measure of dependence between the two variables. To include covariates, we generalize this model to a bivariate Gumbel regression by using $\log(\lambda_I) = \lambda_{I0} + X^\intercal \boldsymbol{\beta}, \log(\lambda_S) = \lambda_{S0} + X^\intercal \boldsymbol{\gamma}$. Under this model, both I and S have marginal exponential distributions with parameters λ_I and λ_S, respectively. Vague priors on all parameters were specified.

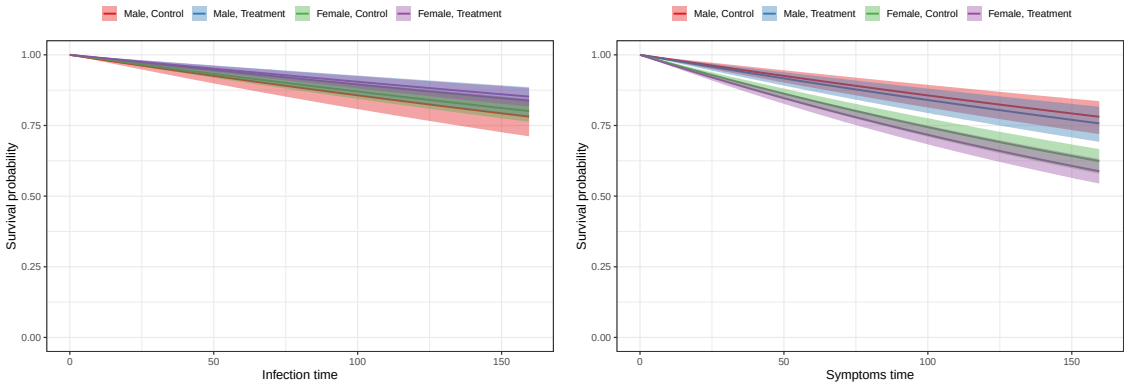


Figure 5.12: Results: Estimated survival curves under the bivariate Gumbel model for infection times (left panel) and times until symptoms (right panel) corresponding to the possible combinations of the binary covariates *gender* and *treatment* fixing the predictor *age* to the average age in the sample.

Some consistent results can be found across the models. For example, under the estimated models women have shorter time until symptoms as measured by the distribution for the corresponding regression coefficient in Figure 5.10 and by the survival curves in Figure 5.11 (right) and Figure 5.12 (right). Unlike inference under the competing models, inference under the proposed bivariate model also shows an effect of the treatment on the infection time. Patients in the intervention group have a delayed re-infection time. The proposed model yields more interpretable results compared to the two independent LDDP models. In fact, under the LDDP models

the probability mass accumulates toward the bounds of the observed censoring times, yielding a "flat" survival curve in the middle region (see Figure 5.11), exactly where we expect events to happen. In fact, most right censored observations are imputed to the right of the rightmost censoring time, whereas most left censored observations are imputed to the left of the leftmost censoring time. This shows that the prior shrinkage alone does not suffice for regularization, and it is consistent with the observations of Section 5.3.1. On the other hand, the Exponential marginal distributions implied by the bivariate Gumbel model represent a very strict parametric assumption that seems not to fit well the data. The specification of such a simple parametric model also yields underestimation of uncertainty. The bivariate Gumbel model also implies a positive correlation between the events, as measured by the 95% credible interval for the parameter α is $[0.72, 0.99]$.

5.8 Discussion

We proposed a novel Bayesian nonparametric bivariate survival regression model that is especially suited for current status data (BNP-CS regression). This research was motivated by the failure of available methods for such data formats. For example, we showed that widely used nonparametric mixture priors lead to biologically uninterpretable results. Our model was built by incorporating simple structural dependence assumptions in a linear dependent Dirichlet process mixture of survival models.

Applied to a recurrent infection study, the method provides novel insights into how symptoms-related hospital visits are affected by covariates. Notably, we

were able to replicate previous results showing a significant effect of the intervention in the randomized clinical trial under consideration. In particular, patients in the intervention group have an improved outlook as measured by delayed reinfections. We also detect an effect of age, with young people having earlier reinfections, which might be due to more risky behaviours. Furthermore, we show that gender has a significant effect on the time until symptoms, but not on infection times. Our study shows that men seek hospital visits later compared to women, suggesting that investing in an awareness campaign could be beneficial.

The ideas presented in this work can be extended to different dependence structures. The present data called for a positive correlation between infection times and infection-related symptom times. A similar model specification can be used for negative correlations. Once the marginal models are flexibly specified, one could for example use copula models to construct a joint distribution with the desired dependence structure. A similar approach, but with positive correlations, could be used for general positively correlated event times when the assumptions used in this application are not available.

Chapter 6

Concluding Remarks

In this chapter we summarize our principal contributions and discuss possible future research directions. A common theme of the three projects presented in this thesis is the development of flexible Bayesian regression methods and computational tools for longitudinal and survival data characterized by varying levels of heterogeneity in parts of their domain.

In Chapter 3, we proposed a novel longitudinal drift-diffusion mixed model for perceptual decision making. Our research was motivated primarily by auditory neuroscience experiments where scientists are interested in understanding how the decision making mechanisms evolve with training. The application to our motivating tone categorization experiments helped formulate interesting novel scientific hypotheses about speech learning. Notably, we discovered that the improvements and the local variations in tone categorization performance can be explained mostly by variations in the parameters reflecting neuron firing rates, while the participants' caution remains constant. We also discovered local groupings among the underlying parameters in various phases of the learning experiments, how they differ between well and poorly performing participants etc. The proposed approach takes the existing literature on drift-diffusion decision making models many steps forward, enabling

neuroscientists to study the longitudinal behavior of biologically interpretable model parameters in much finer detail than what previous methods could achieve.

Methodological extensions and topics of our ongoing research include adapting the proposed models to time constrained learning experiments, accommodating sleep induced overnight 'consolidation' effects, relating our findings with neuroimaging data, etc. Finally, the scope of proposed method is not restricted to auditory neuroscience problems but the approach can be readily applied to study decision making mechanisms in other areas of neuroscience as well.

The crucial element in the development of the drift-diffusion models was to allow the underlying parameters to be similar or different at different longitudinal stages. Such modeling component, which we refer to as local clustering, was initially developed for a single predictor in Chapter 3. Chapter 4 further extends this strategy by developing a flexible Bayesian semiparametric approach to longitudinal functional mixed models in the presence of multiple categorical covariates. Building on mixtures of locally supported B-splines, our proposed method allows the fixed effects components to vary flexibly with the associated covariates, including potentially different sets of covariates at different time points. The mechanism not only allows different sets of covariates to be included in the model at different time points but also allows the selected predictors' influences to vary flexibly over time.

While the focus of Chapter 4 was on dynamically varying partitions for longitudinal data, the same methodology is highly generic and could also be useful in static multiway mixed ANOVA designs. Methodological extensions worth pursuing as topics of separate research include dynamic partition models for the observational

units, as well as spatial and spatiotemporal models for multivariate responses. We also envision to develop principled approaches to accommodate continuous, ordinal and mixed type covariates.

In Chapter 5 we proposed a novel Bayesian nonparametric bivariate survival regression model that is especially suited for current status data. This research was motivated by the failure of available flexible methods to identify the joint distribution of two events from such data formats. In contrast, our model was built by incorporating biologically meaningful dependence assumptions in a linear dependent Dirichlet process mixture of survival models. Applied to a recurrent infection study, our method provides novel insights into how symptoms-related hospital visits are affected by covariates. Moreover, we were able to replicate previous results showing a significant effect of the intervention in the randomized controlled trial under consideration.

The ideas presented in Chapter 5 can be extended to different dependence structures. The motivating study called for a positive correlation between infection times and infection-related symptom times. However, we are currently investigating a copula model specification for outcomes that are negatively correlated, such as overall cancer survival and biomarkers that have a detrimental impact on patients' health.

Bibliography

Agresti, A. (2018). *An introduction to categorical data analysis*. Wiley.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, **88**, 669–679.

Argiento, R., Cremaschi, A., and Vannucci, M. (2020). Hierarchical normalized completely random measures to cluster grouped data. *Journal of the American Statistical Association*, **115**, 318–333.

Barry, D. and Hartigan, J. A. (1992). Product partition models for change point problems. *Annals of Statistics*, **20**, 260–279.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. Springer series in statistics. Springer-Verlag, New York, 2nd edition.

Birdsong, D. (2004). Second language acquisition and ultimate attainment. *Handbook of Applied Linguistics*, pages 82–105.

Bogacz, R., Wagenmakers, E.-J., Forstmann, B. U., and Nieuwenhuis, S. (2010). The neural basis of the speed-accuracy tradeoff. *Trends in Neurosciences*, **33**, 10–16.

Borooah, V. K. (2002). *Logit and probit: Ordered and multinomial models*. Sage.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, **16**, 199–231.

Brody, C. D. and Hanks, T. D. (2016). Neural underpinnings of the evidence accumulator. *Current Opinion in Neurobiology*, **37**, 149–157.

Brown, S. D. and Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, **57**, 153–178.

Brumback, B. A. and Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association*, **93**, 961–976.

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika*, **66**, 429–436.

Cai, B., Lin, X., and Wang, L. (2011). Bayesian proportional hazards model for current status data with monotone splines. *Computational Statistics & Data Analysis*, **55**, 2644–2651.

Camerlenghi, F., Lijoi, A., Orbanz, P., and Prünster, I. (2019). Distribution theory for hierarchical processes. *The Annals of Statistics*, **47**, 67–92.

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in hidden Markov models*. Springer Verlag, Berlin.

Cavanagh, J. F., Wiecki, T. V., Cohen, M. X., Figueroa, C. M., Samanta, J., Sherman, S. J., and Frank, M. J. (2011). Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nature Neuroscience*, **14**, 1462–1467.

Chandra, N. K. and Bhattacharya, S. (2019). Non-marginal decisions: A novel Bayesian multiple testing procedure. *Electronic Journal of Statistics*, **13**, 489–535.

Chandrasekaran, B., Sampath, P. D., and Wong, P. C. (2010). Individual variability in cue-weighting and lexical tone learning. *The Journal of the Acoustical Society of America*, **128**, 456–465.

Chandrasekaran, B., Yi, H.-G., and Maddox, W. T. (2014). Dual-learning systems during speech category learning. *Psychonomic Bulletin & Review*, **21**, 488–495.

Chhikara, R. (1988). *The inverse Gaussian distribution: Theory, methodology, and applications*. CRC Press.

Chib, S. and Hamilton, B. H. (2002). Semiparametric Bayes analysis of longitudinal data treatment models. *Journal of Econometrics*, **110**, 67–89.

Chipman, H. A., George, E. I., McCulloch, R. E., *et al.* (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics*, **4**, 266–298.

Christensen, R. and Johnson, W. (1988). Modelling accelerated failure time with a Dirichlet process. *Biometrika*, **75**, 693–704.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, **34**, 187–202.

Cox, D. R. and Miller, H. D. (1965). *The theory of stochastic processes*. CRC Press.

Craigmile, P. F., Peruggia, M., and Van Zandt, T. (2010). Hierarchical Bayes models for response time data. *Psychometrika*, **75**, 613–632.

Daniels, M. J. and Pourahmadi, M. (2002). Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika*, **89**, 553–566.

de Boor, C. (1978). *A practical guide to splines*. Springer-Verlag.

De Iorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004). An ANOVA model for dependent random measures. *Journal of the American Statistical Association*, **99**, 205–215.

De Iorio, M., Johnson, W. O., Müller, P., and Rosner, G. L. (2009). Bayesian nonparametric nonproportional hazards survival modeling. *Biometrics*, **65**, 762–771.

De Lathauwer, L., De Moore, B., and Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, **21**, 1253–1278.

Denison, D., Adams, N., Holmes, C., and Hand, D. (2002). Bayesian partition modelling. *Computational Statistics & Data Analysis*, **38**, 475–485.

Deshpande, S. K., Bai, R., Balocchi, C., and Starling, J. E. (2020). VC-BART: Bayesian trees for varying coefficients. *arXiv preprint arXiv:2003.06416*.

Díaz-Venegas, C., Downer, B., Langa, K. M., and Wong, R. (2016). Racial and ethnic differences in cognitive function among older adults in the USA. *International Journal of Geriatric Psychiatry*, **31**, 1004–1012.

Diggle, P., Diggle, P. J., Heagerty, P., Heagerty, P. J., Liang, K.-Y., Zeger, S., *et al.* (2002). *Analysis of longitudinal data*. Oxford University Press.

Ding, L. and Gold, J. I. (2013). The basal ganglia's contributions to perceptual decision making. *Neuron*, **79**, 640–649.

Duan, J. A., Guindani, M., and Gelfand, A. E. (2007). Generalized spatial Dirichlet process models. *Biometrika*, **94**, 809–825.

Duan, L. L., Young, A. L., Nishimura, A., and Dunson, D. B. (2020). Bayesian constraint relaxation. *Biometrika*, **107**, 191–204.

Dufau, S., Grainger, J., and Ziegler, J. C. (2012). How to say "no" to a nonword: A leaky competing accumulator model of lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **38**, 1117–1128.

Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society: Series B*, **62**, 355–366.

Dunson, D. B. and Dinse, G. E. (2002). Bayesian models for multivariate current status data with informative censoring. *Biometrics*, **58**, 79–88.

Eddelbuettel, D. and Sanderson, C. (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics & Data Analysis*, **71**, 1054–1063.

Eddelbuettel, D., François, R., Allaire, J., Ushey, K., Kou, Q., Russel, N., Chambers, J., and Bates, D. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, **40**, 1–18.

Efron, B. (2020). Prediction, estimation, and attribution. *Journal of the American Statistical Association*, **115**, 636–655.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89–102.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**, 577–588.

Feng, G., Yi, H. G., and Chandrasekaran, B. (2019). The role of the human auditory corticostriatal network in speech learning. *Cerebral Cortex*, **29**, 4077–4089.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, **1**, 209–230.

Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G. (2008). *Longitudinal data analysis*. CRC Press.

Fontanesi, L., Gluth, S., Spektor, M. S., and Rieskamp, J. (2019). A reinforcement learning diffusion decision model for value-based decisions. *Psychonomic Bulletin & Review*, **26**, 1099–1121.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Science & Business Media.

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, **100**, 1021–1035.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1**, 515–534.

Geweke, J. (1991a). Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities. In *Computing Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface*, pages 571–578.

Geweke, J. (1991b). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Proceedings of the Fourth Valencia International Conference on Bayesian Statistics*, pages 169–193.

Ghahramani, Z. and Jordan, M. I. (1997). Factorial hidden Markov models. *Machine Learning*, **29**, 245–273.

Ghosal, S. and Van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Annals of Statistics*, **27**, 143–158.

Gilbody, S., Littlewood, E., Hewitt, C., Brierley, G., Tharmanathan, P., Araya, R., Barkham, M., Bower, P., Cooper, C., Gask, L., *et al.* (2015). Computerised cognitive behaviour therapy (cCBT) as treatment for depression in primary care (REEACT trial): large scale pragmatic randomised controlled trial. *BMJ*, **351**.

Glimcher, P. W. and Fehr, E. (2013). *Neuroeconomics: Decision making and the brain*. Academic Press.

Gold, J. I. and Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, **30**, 535–574.

Golden, M. R., Whittington, W. L., Handsfield, H. H., Hughes, J. P., Stamm, W. E., Hogben, M., Clark, A., Malinski, C., Helmers, J. R., Thomas, K. K., and Holmes, K. K. (2005). Effect of expedited treatment of sex partners on recurrent or persistent gonorrhea or chlamydial infection. *New England Journal of Medicine*, **352**, 676–685.

Gramacy, R. B., Taddy, M., and Wild, S. M. (2013). Variable selection and sensitivity analysis using dynamic trees, with an application to computer code performance tuning. *Annals of Applied Statistics*, **7**, 51–80.

Groeneboom, P. and Wellner, J. A. (1992). *Information bounds and nonparametric maximum likelihood estimation*, volume 19. Birkhäuser Basel.

Group, C. A. M. P. R. (2000). Long-term effects of budesonide or nedocromil in children with asthma. *New England Journal of Medicine*, **343**, 1054–1063.

Grushka, E. (1972). Characterization of exponentially modified Gaussian peaks in chromatography. *Analytical Chemistry*, **44**, 1733–1738.

Guhaniyogi, R., Qamar, S., and Dunson, D. B. (2017). Bayesian tensor regression. *The Journal of Machine Learning Research*, **18**, 2733–2763.

Gumbel, E. J. (1960). Bivariate exponential distributions. *Journal of the American Statistical Association*, **55**, 698–707.

Guo, W. (2002). Functional mixed effects models. *Biometrics*, **58**, 121–128.

Hanson, T. and Johnson, W. O. (2002). Modeling regression error with a mixture of Polya trees. *Journal of the American Statistical Association*, **97**, 1020–1033.

Hanson, T. and Johnson, W. O. (2004). A Bayesian semiparametric AFT model for interval-censored data. *Journal of Computational and Graphical Statistics*, **13**, 341–361.

Hartigan, J. A. (1990). Partition models. *Communications in Statistics - Theory and Methods*, **19**, 2745–2756.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B*, **55**, 757–796.

Heekeren, H. R., Marrett, S., Bandettini, P. A., and Ungerleider, L. G. (2004). A general mechanism for perceptual decision-making in the human brain. *Nature*, **431**, 859–862.

Hitchcock, F. L. (1927). The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, **6**, 164–189.

Hjort, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *The Annals of Statistics*, **18**, 1259–1294.

Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, **85**, 809–822.

Hothorn, T. and Everitt, B. S. (2014). *A handbook of statistical analyses using R*. CRC press.

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *The Annals of Statistics*, **24**, 540–568.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, **96**, 161–173.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., and Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, **87**, 47–57.

James, L. F., Lijoi, A., and Prünster, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics*, **36**, 76–97.

Jara, A., Lesaffre, E., De Iorio, M., and Quintana, F. (2010). Bayesian semiparametric inference for multivariate doubly-interval-censored data. *The Annals of Applied Statistics*, **4**, 2126–2149.

Jara, A., Hanson, T., Quintana, F., Müller, P., and Rosner, G. (2011). DPpackage: Bayesian semi- and nonparametric modeling in R. *Journal of Statistical Software*, **40**, 1–30.

Jewell, N. P. and Shiboski, S. C. (1990). Statistical analysis of HIV infectivity based on partner studies. *Biometrics*, **46**, 1133–1150.

Johndrow, J. E., Bhattacharya, A., and Dunson, D. B. (2017). Tensor decompositions and sparse log-linear models. *Annals of Statistics*, **45**, 1–38.

Johnson, J. S. and Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, **21**, 60–99.

Kalbfleisch, J. D. (1978). Non-parametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society: Series B*, **40**, 214–221.

Kalli, M., Griffin, J. E., and Walker, S. G. (2011). Slice sampling mixture models. *Statistics and Computing*, **21**, 93–105.

Kim, S., Potter, K., Craigmile, P. F., Peruggia, M., and Van Zandt, T. (2017). A Bayesian race model for recognition memory. *Journal of the American Statistical Association*, **112**, 77–91.

Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review*, **51**, 455–500.

Koslovsky, M. D., Hébert, E. T., Businelle, M. S., Vannucci, M., *et al.* (2020). A Bayesian time-varying effect model for behavioral mHealth data. *Annals of Applied Statistics*, **14**, 1878–1902.

Kottas, A. and Gelfand, A. E. (2001). Bayesian semiparametric median regression modeling. *Journal of the American Statistical Association*, **96**, 1458–1468.

Kunkel, D., Potter, K., Craigmile, P. F., Peruggia, M., and Van Zandt, T. (2019). A Bayesian race model for response times under cyclic stimulus discriminability. *The Annals of Applied Statistics*, **13**, 271–296.

Kuo, L. and Mallick, B. (1997). Bayesian semiparametric inference for the accelerated failure-time model. *Canadian Journal of Statistics*, **25**, 457–472.

Leite, F. P. and Ratcliff, R. (2010). Modeling reaction time and accuracy of multiple-alternative decisions. *Attention, Perception, & Psychophysics*, **72**, 246–273.

Li, Y., Lin, X., and Müller, P. (2010). Bayesian inference in semiparametric mixed models for longitudinal data. *Biometrics*, **66**, 70–78.

Linero, A. R. and Yang, Y. (2018). Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society: Series B*, **80**, 1087–1110.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics*, **12**, 351–357.

Lu, J. (1995). *Degradation processes and related reliability models*. Ph.D. thesis, McGill University, Montreal, Canada.

MacEachern, S. N. (1999). Dependent nonparametric processes. In *ASA proceedings of the section on Bayesian statistical science*, volume 1.

MacEachern, S. N. and Müller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, **7**, 223–238.

MacLehose, R. F. and Dunson, D. B. (2009). Nonparametric Bayes kernel-based priors for functional data analysis. *Statistica Sinica*, **19**, 611–629.

Maddox, W. T. and Chandrasekaran, B. (2014). Tests of a dual-system model of speech category learning. *Bilingualism: Language and Cognition*, **17**, 709–728.

Makalic, E. and Schmidt, D. F. (2016). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, **23**, 179–182.

McDonald, S. and Zucchini, W. (1997). *Hidden Markov and other models for discrete-valued time series*. Chapman & Hall, London.

Milosavljevic, M., Malmaud, J., Huth, A., Koch, C., and Rangel, A. (2010). The drift diffusion model can account for the accuracy and reaction time of value-based choices under high and low time pressure. *Judgment and Decision Making*, **5**, 437–449.

Moore, W., Pedlow, S., Krishnamurty, P., and Wolter, K. (2000). National longitudinal survey of youth 1997 (NLSY97). Technical report, National Opinion Research Center.

Morris, J. S. (2015). Functional regression. *Annual Review of Statistics and Its Application*, **2**, 321–359.

Morris, J. S. and Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society: Series B*, **68**, 179–199.

Müller, P., Parmigiani, G., Robert, C., and Rousseau, J. (2004). Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American Statistical Association*, **99**, 990–1001.

Müller, P., Quintana, F. A., Rosner, G. L., and Maitland, M. L. (2013). Bayesian inference for longitudinal data with non-parametric treatment effects. *Biostatistics*, **15**, 341–352.

Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015). *Bayesian nonparametric data analysis*. Springer.

Navarro, D. J. and Fuss, I. G. (2009). Fast and accurate calculations for first-passage times in Wiener diffusion models. *Journal of Mathematical Psychology*, **53**, 222–230.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, **9**, 249–265.

Neal, R. M. (2003). Slice sampling. *Annals of Statistics*, **31**, 705–767.

Nguyen, X. (2010). Inference of global clusters from locally distributed data. *Bayesian Analysis*, **5**, 817–845.

Nguyen, X. and Gelfand, A. E. (2011). The Dirichlet labeling process for clustering functional data. *Statistica Sinica*, **21**, 1249–1289.

Nguyen, X. and Gelfand, A. E. (2014). Bayesian nonparametric modeling for functional analysis of variance. *Annals of the Institute of Statistical Mathematics*, **66**, 495–526.

Page, G. L., Quintana, F. A., and Dahl, D. B. (2020). Dependent random partition models. *arXiv preprint arxiv:1912.11542*.

Papadogeorgou, G., Zhang, Z., and Dunson, D. B. (2019). Soft tensor regression. *arXiv preprint arXiv:1910.09699*.

Patra, S. and Dunson, D. B. (2018). Constrained Bayesian inference through posterior projections. *arXiv preprint arXiv:1812.05741*.

Paulon, G., Reetzke, R., Chandrasekaran, B., and Sarkar, A. (2019). Functional logistic mixed-effects models for learning curves from longitudinal binary data. *Journal of Speech, Language, and Hearing Research*, **62**, 543–553.

Paulon, G., Llanos, F., Chandrasekaran, B., and Sarkar, A. (2020). Bayesian semiparametric longitudinal drift-diffusion mixed models for tone learning in adults. *Journal of the American Statistical Association*, pages 1–14.

Pedersen, M. L., Frank, M. J., and Biele, G. (2017). The drift diffusion model as the choice rule in reinforcement learning. *Psychonomic Bulletin & Review*, **24**, 1234–1251.

Peters, J. and D'Esposito, M. (2020). The drift diffusion model as the choice rule in inter-temporal and risky choice: A case study in medial orbitofrontal cortex lesion patients and controls. *PLOS Computational Biology*, **16**.

Petrone, S., Guindani, M., and Gelfand, A. E. (2009). Hybrid Dirichlet mixture models for functional data. *Journal of the Royal Statistical Society: Series B*, **71**, 755–782.

Polson, N. G. and Scott, J. G. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, **7**, 887–902.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, **108**, 1339–1349.

Proudfoot, J., Goldberg, D., Mann, A., Everitt, B., Marks, I., and Gray, J. (2003). Computerized, interactive, multimedia cognitive-behavioural program for anxiety and depression in general practice. *Psychological Medicine*, **33**, 217–227.

Purcell, B. A. (2013). *Neural mechanisms of perceptual decision making*. Vanderbilt University.

Quintana, F. A. and Iglesias, P. L. (2003). Bayesian clustering and product partition models. *Journal of the Royal Statistical Society: Series B*, **65**, 557–574.

Quintana, F. A., Johnson, W. O., Waetjen, L. E., and B. Gold, E. (2016). Bayesian nonparametric longitudinal data analysis. *Journal of the American Statistical Association*, **111**, 1168–1181.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *IEEE*, **77**, 257–286.

Ramsay, J. O. and Silverman, B. W. (2007). *Applied functional data analysis: Methods and case studies*. Springer.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, **85**, 59–108.

Ratcliff, R. and McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, **20**, 873–922.

Ratcliff, R. and Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, **9**, 347–356.

Ratcliff, R., Smith, P. L., Brown, S. D., and McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, **20**, 260–281.

Reetzke, R., Xie, Z., Llanos, F., and Chandrasekaran, B. (2018). Tracing the trajectory of sensory plasticity across different stages of speech learning in adulthood. *Current Biology*, **28**, 1419–1427.

Roark, C. L., Paulon, G., Sarkar, A., and Chandrasekaran, B. (2021). Comparing perceptual category learning across modalities in the same individuals. *Psychonomic Bulletin & Review*, **28**, 898–909.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, **18**, 349–367.

Roberts, G. O., Rosenthal, J. S., *et al.* (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, **16**, 351–367.

Rodriguez, A., Dunson, D. B., and Gelfand, A. E. (2008). The nested Dirichlet process. *Journal of the American Statistical Association*, **103**, 1131–1154.

Ross, S. M., Kelly, J. J., Sullivan, R. J., Perry, W. J., Mercer, D., Davis, R. M., Washburn, T. D., Sager, E. V., Boyce, J. B., and Bristow, V. L. (1996). *Stochastic processes*. Wiley New York.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, **11**, 735–757.

Sal y Rosas, V. G. and Hughes, J. P. (2011). Nonparametric and semiparametric analysis of current status data subject to outcome misclassification. *Statistical Communications in Infectious Diseases*, **3**, 364.

Sarkar, A. and Dunson, D. B. (2016). Bayesian nonparametric modeling of higher order Markov chains. *Journal of the American Statistical Association*, **111**, 1791–1803.

Schall, J. D. (2001). Neural basis of deciding, choosing and acting. *Nature Reviews Neuroscience*, **2**, 33–42.

Scott, S. L. (2002). Bayesian methods for hidden Markov models recursive computing in the 21st century. *Journal of the American Statistical Association*, **97**, 337–351.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639–650.

Singer, J. D., Willett, J. B., Willett, J. B., *et al.* (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press.

Singmann, H., Brown, S., Gretton, M., and Heathcote, A. (2019). rtdists: Response time distributions. R package version 0.10-0.

Smith, P. L. and Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, **27**, 161–168.

Smith, P. L. and Vickers, D. (1988). The accumulator model of two-choice discrimination. *Journal of Mathematical Psychology*, **32**, 135–168.

Song, J. H., Skoe, E., Wong, P. C., and Kraus, N. (2008). Plasticity in the adult human auditory brainstem following short-term linguistic training. *Journal of Cognitive Neuroscience*, **20**, 1892–1902.

Sonnega, A., Faul, J. D., Ofstedal, M. B., Langa, K. M., Phillips, J. W., and Weir, D. R. (2014). Cohort profile: The health and retirement study (HRS). *International Journal of Epidemiology*, **43**, 576–585.

Sparapani, R. A., Logan, B. R., McCulloch, R. E., and Laud, P. W. (2016). Nonparametric survival analysis using Bayesian additive regression trees (BART). *Statistics in Medicine*, **35**, 2741–2753.

Starling, J. E., Murray, J. S., Carvalho, C. M., Bukowski, R. K., and Scott, J. G. (2020). BART with targeted smoothing: An analysis of patient-specific stillbirth risk. *Annals of Applied Statistics*, **14**, 28–50.

Suarez, A. J. and Ghosal, S. (2016). Bayesian clustering of functional data using local features. *Bayesian Analysis*, **11**, 71–98.

Taddy, M. A., Gramacy, R. B., and Polson, N. G. (2011). Dynamic trees for learning and design. *Journal of the American Statistical Association*, **106**, 109–123.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2005). Sharing clusters among related groups: Hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems*, pages 1385–1392.

Teichert, T., Grinband, J., and Ferrera, V. (2016). The importance of decision onset. *Journal of Neurophysiology*, **115**, 643–661.

Titsias, M. K. and Yau, C. (2014). Hamming ball auxiliary sampling for factorial hidden Markov models. In *Advances in Neural Information Processing Systems*, pages 2960–2968.

Tucker, L. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrica*, **31**, 273–282.

Tuerlinckx, F. (2004). The efficient computation of the cumulative distribution and probability density functions in the diffusion model. *Behavior Research Methods, Instruments, & Computers*, **36**, 702–716.

Tuerlinckx, F., Maris, E., Ratcliff, R., and De Boeck, P. (2001). A comparison of four methods for simulating the diffusion process. *Behavior Research Methods, Instruments, & Computers*, **33**, 443–456.

Usher, M. and McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, **108**, 550–592.

Van der Groen, O., Tang, M. F., Wenderoth, N., and Mattingley, J. B. (2018). Stochastic resonance enhances the rate of evidence accumulation during combined brain stimulation and perceptual decision-making. *PLOS Computational Biology*, **14**, 1–17.

Van Gael, J., Saatci, Y., Teh, Y. W., and Ghahramani, Z. (2008). Beam sampling for the infinite hidden Markov model. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1088–1095. ACM.

Vandekerckhove, J. and Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review*, **14**, 1011–1026.

Vandekerckhove, J., Tuerlinckx, F., and Lee, M. D. (2008). A Bayesian approach to diffusion process models of decision-making. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages 1429–1434. Washington, DC.

Wade, S. and Ghahramani, Z. (2018). Bayesian cluster analysis: Point estimation and credible balls. *Bayesian Analysis*, **13**, 559–626.

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, **3**, 257–295.

Wang, N., Wang, L., and McMahan, C. S. (2015). Regression analysis of bivariate current status data under the Gamma-frailty proportional hazards model using the EM algorithm. *Computational Statistics & Data Analysis*, **83**, 140–150.

Wang, W. and Ding, A. A. (2000). On assessing the association for bivariate current status data. *Biometrika*, **87**, 879–893.

Wang, Y., Spence, M. M., Jongman, A., and Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America*, **106**, 3649–3658.

Whitmore, G. and Seshadri, V. (1987). A heuristic derivation of the inverse gaussian distribution. *The American Statistician*, **41**, 280–281.

Wilson, R. S., Capuano, A. W., Sytsma, J., Bennett, D. A., and Barnes, L. L. (2015). Cognitive aging in older black and white persons. *Psychology and Aging*, **30**, 279–285.

Wong, P. C., Vuong, L. C., and Liu, K. (2017). Personalized learning: From neurogenetics of behaviors to designing optimal language training. *Neuropsychologia*, **98**, 192–200.

Xie, Z., Reetzke, R., and Chandrasekaran, B. (2017). Stability and plasticity in neural encoding of linguistically relevant pitch patterns. *Journal of Neurophysiology*, **117**, 1409–1424.

Xue, H., Lam, K., and Li, G. (2004). Sieve maximum likelihood estimator for semiparametric regression models with current status data. *Journal of the American Statistical Association*, **99**, 346–356.

Zanella, G. (2019). Informed proposals for local MCMC in discrete spaces. *Journal of the American Statistical Association*, pages 1–14.

Zucchini, W., MacDonald, I. L., and Langrock, R. (2017). *Hidden Markov models for time series: An introduction using R*. CRC press.